FALSE DISCOVERIES OCCUR EARLY ON THE LASSO PATH

By

Weijie Su
Małgorzata Bogdan
Emmanuel Candès

Technical Report No. 2015-20
November 2015

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

FALSE DISCOVERIES OCCUR EARLY ON THE LASSO PATH

By

Weijie Su
Stanford University

Małgorzata Bogdan
Wroclaw University of Technology,
Poland

Emmanuel Candès
Stanford University

Technical Report No. 2015-20
November 2015

**This research was supported in part by
National Science Foundation grant CCF 0963835.**

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

http://statistics.stanford.edu

False Discoveries Occur Early on the Lasso Path

Weijie Su* Małgorzata Bogdan†,‡ Emmanuel Candès*

November 5, 2015

* Department of Statistics, Stanford University, Stanford, CA 94305, USA
† Faculty of Mathematics, Wroclaw University of Technology, Poland
‡ Institute of Mathematics, University of Wroclaw, Poland

Abstract

In regression settings where explanatory variables have very low correlations and where there are relatively few effects each of large magnitude, it is commonly believed that the Lasso shall be able to find the important variables with few errors—if any. In contrast, this paper shows that this is not the case even when the design variables are stochastically independent. In a regime of linear sparsity, we demonstrate that true features and null features are always interspersed on the Lasso path, and that this phenomenon occurs no matter how strong the effect sizes are. We derive a sharp asymptotic trade-off between false and true positive rates or, equivalently, between measures of type I and type II errors along the Lasso path. This trade-off states that if we ever want to achieve a type II error (false negative rate) under a given threshold, then anywhere on the Lasso path the type I error (false positive rate) will need to exceed a given threshold so that we can never have both errors at a low level at the same time. Our analysis uses tools from approximate message passing (AMP) theory as well as novel elements to deal with a possibly adaptive selection of the Lasso regularizing parameter.

Keywords. Lasso, Lasso path, false discovery rate, false negative rate, power, approximate message passing, adaptive selection of parameters.

1 Introduction

Almost all data scientists know about and routinely use the Lasso [25, 27] to fit regression models. In the big data era, where the number p of explanatory variables often exceeds the number n of observational units, it may even supersede the method of least-squares. One appealing feature of the Lasso over earlier techniques such as ridge regression is that it automatically performs variable reduction, since it produces models where lots of—if not most—regression coefficients are estimated to be exactly zero. In high-dimensional problems where p is either comparable to n or even much larger, the belief is that the Lasso will select those important variables out of a sea of potentially many irrelevant features.

Imagine we have an $n \times p$ design matrix $\boldsymbol{X}$ of features, and an n-dimensional response $\boldsymbol{y}$ obeying the standard linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{z},$$

where $\boldsymbol{z}$ is a noise term. The Lasso is the solution to

$$\widehat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{b}\in\mathbb{R}^p}{\operatorname{argmin}} \ \tfrac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2 + \lambda \|\boldsymbol{b}\|_1; \tag{1.1}$$

if we think of the noise term as being Gaussian, we interpret it as a penalized maximum likelihood estimate, in which the fitted coefficients are penalized in an ℓ_1 sense, thereby encouraging sparsity. (There are nowadays many variants on this idea including ℓ_1-penalized logistic regression [27], elastic nets [36], graphical Lasso [32], adaptive Lasso [35], and many others.) As is clear from (1.1), the Lasso depends upon a regularizing parameter λ, which must be chosen in some fashion: in a great number of applications this is typically done via adaptive or data-driven methods; for instance, by cross-validation [13, 18, 33, 23]. Below, we will refer to the Lasso path as the family of solutions $\widehat{\boldsymbol{\beta}}(\lambda)$ as λ varies between 0 and ∞. A variable j is selected at λ if $\widehat{\beta}_j(\lambda) \neq 0$.[1]

The Lasso is, of course, mostly used in situations where the true regression coefficient sequence is suspected to be sparse or nearly sparse. In such settings, researchers often believe—or, at least, wish—that as long as the true signals (the nonzero regression coefficients) are sufficiently strong compared to the noise level, the Lasso with a carefully tuned value of λ will select most of the true signals while picking out very few, if any, noise variables. This faith has served as a popular assumption, especially in the selective inference literature, see [19] and its many follow-ups; there, one studies the distributions of statistics under the assumption that all variables truly in the model (with $\beta_j \neq 0$) have been selected before any—or very few—'null' variable (with $\beta_j = 0$). However, the major result of this paper is that in a regime of *linear sparsity* defined below, this can never be the case—at least under random designs. *We rigorously demonstrate that true features and null features are always interspersed on the Lasso path, and that this phenomenon occurs no matter how strong the signals are.*

Formally, we derive a fundamental trade-off between power (the ability to detect signals) and type I errors or, said differently, between the true positive and the false positive rates. This trade-off says that it is impossible to achieve high power and a low false positive rate simultaneously. Formally, letting the false discovery proportion (FDP) and the true positive proportion (TPP) be defined as usual (see (2.1) and (2.2)), we compute the formula for an exact boundary curve separating achievable $(\mathrm{TPP}, \mathrm{FDP})$ pairs from pairs that are impossible to achieve no matter the value of the signal-to-noise ratio (SNR). Hence, there is a whole favorable region in the $(\mathrm{TPP}, \mathrm{FDP})$ plane that cannot be reached, see Figure 3 for an illustration. A different way to phrase the trade-off is via false discovery and false negative rates. Here, the FDP is a natural measure of type I error while $1 - \mathrm{TPP}$ (often called the false negative proportion) is the fraction of missed signals, a natural notion of type II error. In this language, our results simply say that nowhere on the Lasso path can both types of error rates be simultaneously low.

As a setup for our theoretical findings, Figure 1 compares the performance of the Lasso under a 1000×1000 orthogonal design, and under a random Gaussian design of the same size. In the latter setting, the entries of $\boldsymbol{X}$ are independent draws from $\mathcal{N}(0, 1/1000)$ so that each column is approximately normalized. Here, note that $n = p$ so that we are not even really discussing a high-dimensional scenario. Set $\beta_1 = \cdots = \beta_{200} = 50$, $\beta_{201} = \cdots = \beta_{1000} = 0$ and the errors to be independent standard normals. Hence, we have 200 nonzero coefficients out of 1000 (a relatively

[1] We also say that a variable j enters the Lasso path at λ_0 if there is there is $\varepsilon > 0$ such that $\widehat{\beta}_j(\lambda) = 0$ for $\lambda \in [\lambda_0 - \varepsilon, \lambda_0]$ and $\widehat{\beta}_j(\lambda) \neq 0$ for $\lambda \in (\lambda_0, \lambda_0 + \varepsilon]$. Similarly a variable is dropped at λ_0 if $\widehat{\beta}_j(\lambda) \neq 0$ for $\lambda \in [\lambda_0 - \varepsilon, \lambda_0)$ and $\widehat{\beta}_j(\lambda) = 0$ for $\lambda \in [\lambda_0, \lambda_0 + \varepsilon]$.

sparse setting), all 50 times higher than the noise level in magnitude (very large SNR). In the orthogonal design, all the true predictors enter the Lasso path first, as expected. This is in sharp contrast with the random design setting, in which the Lasso selects null variables rather early. When the Lasso includes half of the true predictors so that the false negative proportion falls below 50%, the FDP has already passed 10% meaning that we have already made 11 false discoveries. The FDP further increases to 18% the first time the Lasso model includes all true predictors, i.e. achieves full power (false negative proportion vanishes).

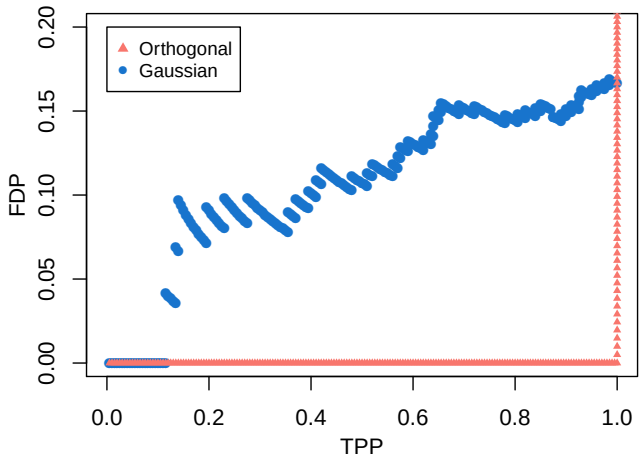


Figure 1: True positive and false positive rates along the Lasso path under the orthogonal and Gaussian random designs.

Figure 2 provides a closer look at this phenomenon, and summarizes the outcomes from 100 independent experiments under the same Gaussian random design setting. In all the simulations, the first noise variable enters the Lasso model before 40% of the true signals are detected, and the last true signal is preceded by at least 21 and, sometimes, even 64 false discoveries. On average, the Lasso detects about 32 signals before the first false variable enters; to put it differently, the TPP is only 16.4% at the time the first false discovery is made. The average FDP evaluated the first time when all signals are detected is 14.0%.

2 The Lasso Trade-off Diagram

Our theoretical results rigorously confirm that the observed Lasso behavior for model selection holds under certain assumptions. We mostly work in the setting of [3], which specifies the design $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, the parameter sequence $\boldsymbol{\beta} \in \mathbb{R}^p$ and the errors $\boldsymbol{z} \in \mathbb{R}^n$.

Working hypothesis The design matrix $\boldsymbol{X}$ has i.i.d. $\mathcal{N}(0, 1/n)$ entries and the errors z_i are i.i.d. $\mathcal{N}(0, \sigma^2)$, where σ is fixed but otherwise arbitrary. Note that we do not exclude the value $\sigma = 0$ corresponding to noiseless observations. The regression coefficients $\beta_1, \ldots, \beta_p$ are independent copies of a random variable Π obeying $\mathbb{E}\,\Pi^2 < \infty$ and $\mathbb{P}(\Pi \neq 0) = \epsilon \in (0, 1)$ for some constant ϵ. For completeness, $\boldsymbol{X}, \boldsymbol{\beta}$, and $\boldsymbol{z}$ are all independent from each other. As in [3], we are interested in the limiting case where $p, n \to \infty$ with $n/p \to \delta$ for some positive constant δ. A few comments are in order.

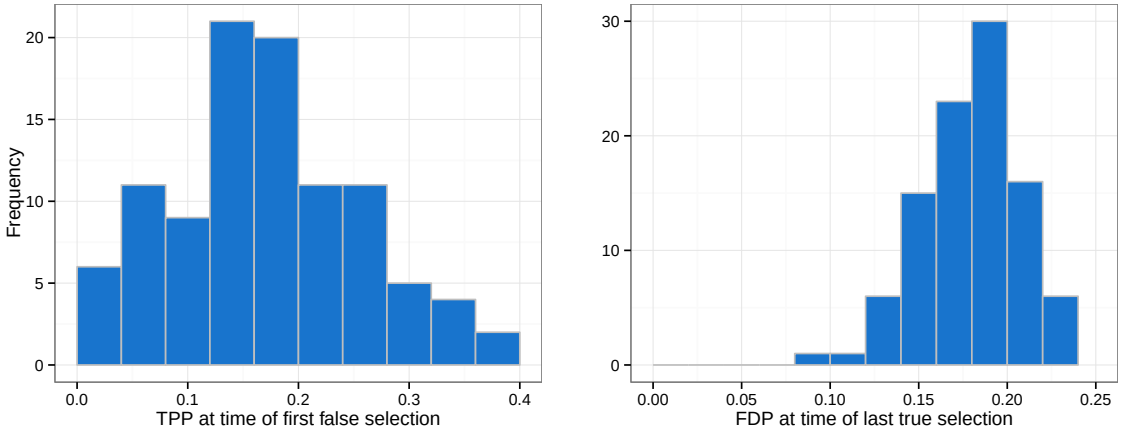


Figure 2: Left: power when the first false variable enters the Lasso model. Right: false discovery proportion the first time power reaches one (false negative proportion vanishes).

- The first concerns the degree of sparsity. In our model, the expected number of nonzero regression coefficients is linear in p and equal to $\epsilon \cdot p$ for some $\epsilon > 0$. Hence, this model excludes a form of asymptotics, where the fraction of nonzero coefficients vanishes in the limit of large problem sizes.
- Second, Gaussian designs with independent columns are believed to be "easy" or favorable for model selection due to weak correlations between distinct features. (Such designs happen to obey restricted isometry properties [6] or restricted eigenvalue conditions [4] with high probability, which have been shown to be useful in settings sparser than those considered in this paper.) Hence, negative results under the working hypothesis are likely to extend more generally.
- Third, the assumption concerning the distribution of the regression coefficients can be slightly weakened: all we need is that the sequence $\beta_1, \ldots, \beta_p$ has a convergent empirical distribution with bounded second moment. We shall not pursue this generalization here.

2.1 Main result

Throughout the paper, V (resp. T) denotes the number of Lasso false (resp. true) discoveries while $k = |\{j : \beta_j \neq 0\}|$ denotes the number of true signals; formally, $V(\lambda) = |\{j : \widehat{\beta}_j(\lambda) \neq 0 \text{ and } \beta_j = 0\}|$ whereas $T(\lambda) = |\{j : \widehat{\beta}_j(\lambda) \neq 0 \text{ and } \beta_j \neq 0\}|$. With this, we define the FDP as usual ($a \vee b = \max\{a, b\}$),

$$\mathrm{FDP}(\lambda) = \frac{V(\lambda)}{|\{j : \widehat{\beta}_j(\lambda) \neq 0\}| \vee 1} \tag{2.1}$$

and, similarly, the TPP is defined as

$$\mathrm{TPP}(\lambda) = \frac{T(\lambda)}{k \vee 1}. \tag{2.2}$$

The dependency on λ shall often be suppressed when clear from the context. Our main result provides an explicit trade-off between FDP and TPP.

Theorem 1. *Fix $\delta \in (0, \infty)$ and $\epsilon \in (0, 1)$, and consider the function $q^\star(\cdot) = q^\star(\cdot; \delta, \epsilon) > 0$ given in (2.4). Then under the working hypothesis, the following conclusions hold:*

(a) In the **noiseless case** *($\sigma = 0$), the event*

$$\bigcap_{\lambda \geq 0.01} \Big\{ \mathrm{FDP}(\lambda) \geq q^\star\left(\mathrm{TPP}(\lambda)\right) - 0.001 \Big\} \tag{2.3}$$

holds with probability tending to one. (The lower bound on λ in (2.3) does not impede interpretability since we are not interested in variables entering the path last.)

(b) With **noisy data** *($\sigma > 0$) the conclusion is exactly the same as in (a).*

(c) Therefore, in both the noiseless and noisy cases, no matter how we choose $\widehat{\lambda}(\boldsymbol{y}, \boldsymbol{X}) \geq 0.01$ **adaptively** *by looking at the response $\boldsymbol{y}$ and design $\boldsymbol{X}$, with probability tending to one we will never have $\mathrm{FDP}(\widehat{\lambda}) < q^\star(\mathrm{TPP}(\widehat{\lambda})) - 0.001$.*

(d) The boundary curve $q^\star$ is **tight***: any continuous curve $q(u) \geq q^\star(u)$ with strict inequality for some u will fail (a) and (b) for some prior distribution Π on the regression coefficients.*

The numerical values 0.01 and 0.001 in (2.3) are arbitrary, and can be replaced by any positive numerical constants.

We would like to emphasize that the boundary is derived from a best-case point of view. For a fixed prior Π, we also provide in Theorem 3 from Appendix C a trade-off curve q^Π between TPP and FDP, which always lies above the boundary $q^\star$. Hence, the trade-off is of course less favorable when we deal with a specific Lasso problem. In fact, $q^\star$ is nothing else but the lower envelope of all the instance-specific curves q^Π with $\mathbb{P}(\Pi \neq 0) = \epsilon$.

Figure 3 presents two instances of the *Lasso trade-off diagram*, where the curve $q^\star(\cdot)$ separates the red region, where both type I and type II errors are small, from the rest (the white region). Looking at this picture, Theorem 1 says that nowhere on the Lasso path we will find ourselves in the red region, and that this statement continues to hold true even when there is no noise. Our theorem also says that we cannot move the boundary upward. As we shall see, we can come arbitrarily close to any point on the curve by specifying a prior Π and a value of λ. Note that the right plot is vertically truncated at 0.6791, implying that TPP cannot even approach 1 in the regime of $\delta = 0.3, \epsilon = 0.15$. This upper limit is where the Donoho-Tanner phase transition occurs [12], see the discussion in Appendix B.

Support recovery from noiseless data is presumably the most ideal scenario. Yet, the trade-off remains the same as seen in the first claim of the theorem. As explained in Section 3, this can be understood by considering that the root cause underlying the trade-off in both the noiseless and noisy cases come from the pseudo-noise introduced by shrinkage.

We now turn to specify $q^\star$. For a fixed u, let $t^\star(u)$ be the largest positive root[2] of the equation in t,

$$\frac{2(1-\epsilon)\left[(1+t^2)\Phi(-t) - t\phi(t)\right] + \epsilon(1+t^2) - \delta}{\epsilon\left[(1+t^2)(1-2\Phi(-t)) + 2t\phi(t)\right]} = \frac{1-u}{1-2\Phi(-t)}.$$

[2]If $u = 0$, treat $+\infty$ as a root of the equation, and in (2.4) conventionally set $0/0 = 0$. In the case where $\delta \geq 1$, or $\delta < 1$ and ϵ is no larger than a threshold determined only by δ, the range of u is the unit interval $[0, 1]$. Otherwise, the range of u is the interval with endpoints 0 and some number strictly smaller than 1, see the discussion in Appendix B.

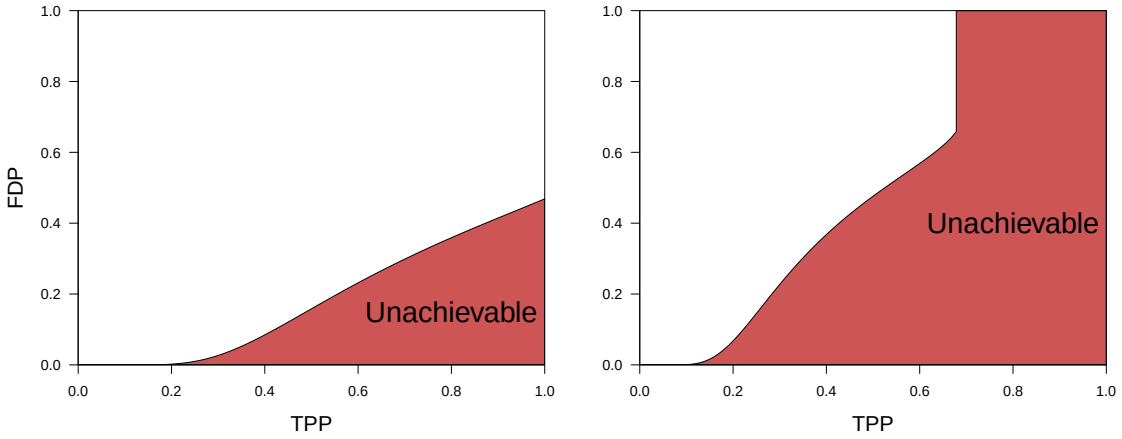


Figure 3: The Lasso trade-off diagram: left is with $\delta = 0.5$ and $\epsilon = 0.15$, and right is with $\delta = 0.3$ and $\epsilon = 0.15$ (the vertical truncation occurs at 0.6791).

Then

$$q^\star(u; \delta, \epsilon) = \frac{2(1-\epsilon)\Phi(-t^\star(u))}{2(1-\epsilon)\Phi(-t^\star(u)) + \epsilon u}. \tag{2.4}$$

It can be shown that this function is infinitely many times differentiable over its domain, always strictly increasing, and vanishes at $u = 0$. Matlab code to calculate $q^\star$ is available at `http://wjsu.web.stanford.edu/code.html`.

Figure 4 displays examples of the function $q^\star$ for different values of ϵ (sparsity), and δ (dimensionality). It can be observed that the issue of FDR control becomes more severe when the sparsity ratio $\epsilon = k/p$ increases and the dimensionality $1/\delta = p/n$ increases.

2.2 Numerical illustration

Figure 5 provides the outcomes of numerical simulations for finite values of n and p in the noiseless setup where $\sigma = 0$. For each of $n = p = 1000$ and $n = p = 5000$, we compute 10 independent Lasso paths and plot all pairs $(\mathrm{TPP}, \mathrm{FDP})$ along the way. In Figure 5a we can see that when $\mathrm{TPP} < 0.8$, then the large majority of pairs $(\mathrm{TPP}, \mathrm{FDP})$ along these 10 paths are above the boundary. When TPP approaches one, the average FDP becomes closer to the boundary and a fraction of the paths fall below the line. As expected this proportion is substantially smaller for the larger problem size.

2.3 Sharpness

The last conclusion from the theorem stems from the following fact: take any point $(u, q^\star(u))$ on the boundary curve; then we can approach this point by fixing $\epsilon' \in (0, 1)$ and setting the prior to be

$$\Pi = \begin{cases} M, & \text{w.p. } \epsilon \cdot \epsilon', \\ M^{-1}, & \text{w.p. } \epsilon \cdot (1-\epsilon'), \\ 0, & \text{w.p. } 1-\epsilon. \end{cases}$$

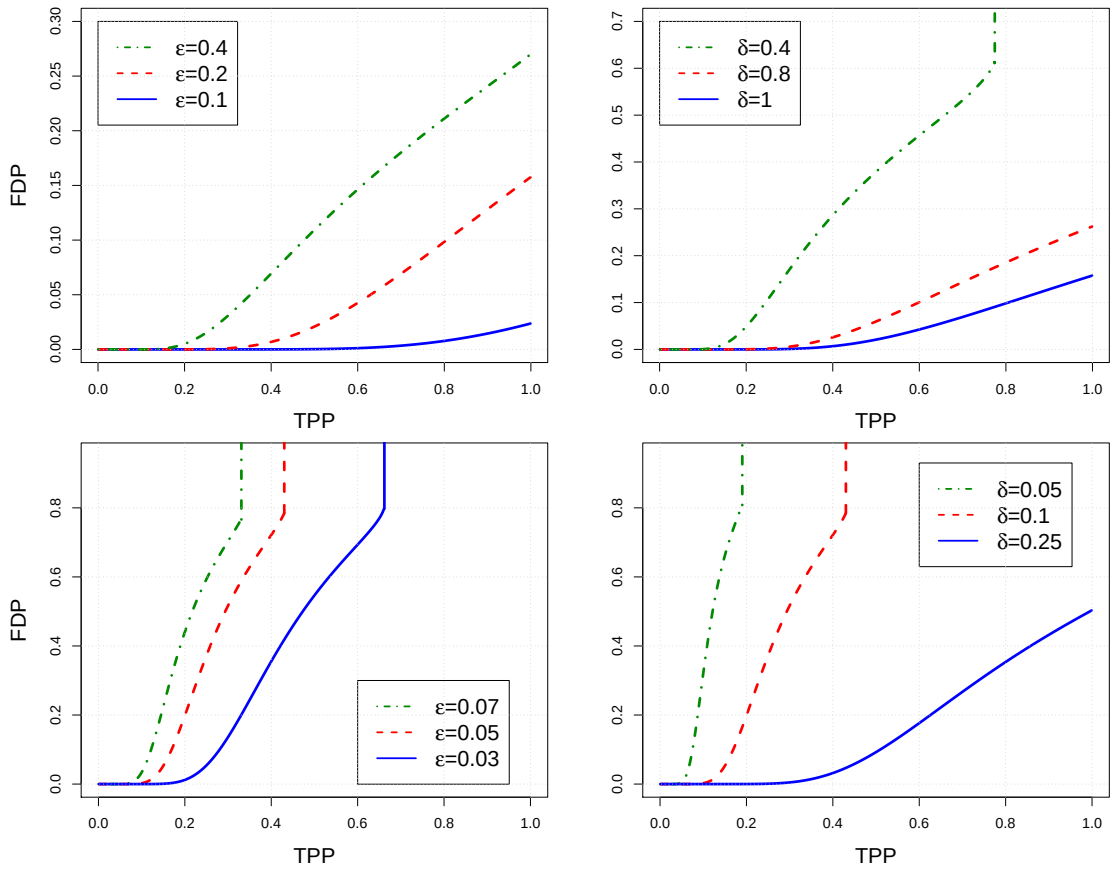
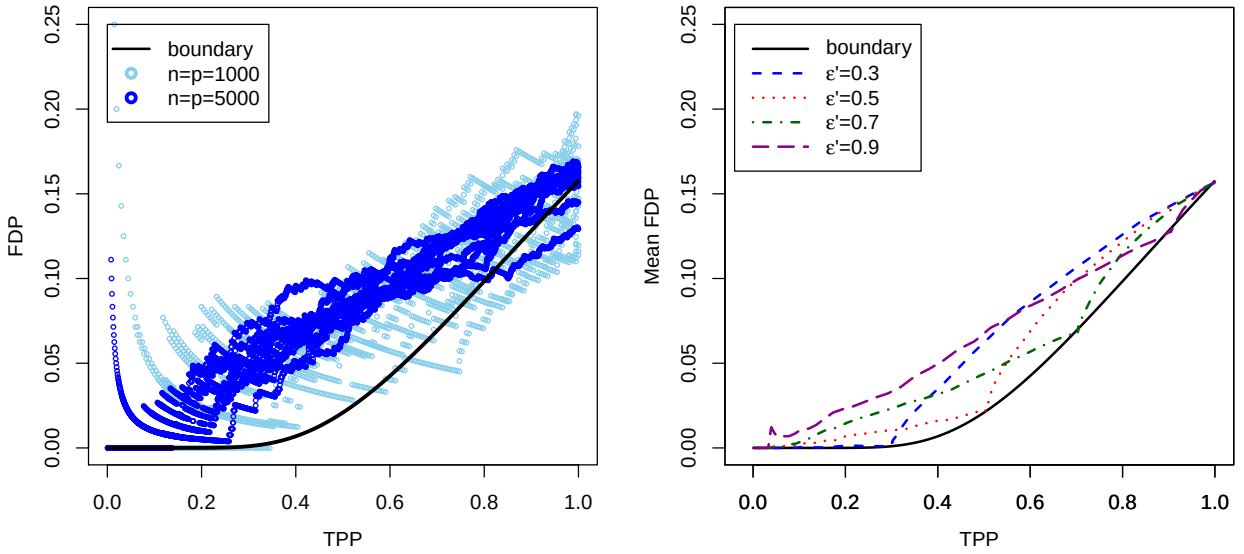


Figure 4: Top-left is with $\delta = 1$; top-right is with $\epsilon = 0.2$; bottom-left is with $\delta = 0.1$; and bottom-right is with $\epsilon = 0.05$.



(a) Realized $(\mathrm{TPP}, \mathrm{FDP})$ pairs

(b) Sharpness of the boundary

Figure 5: In both (a) and (b), $n/p = \delta = 1$, $\epsilon = 0.2$, and the noise level is $\sigma = 0$ (noiseless). (a) FDP vs. TPP along 10 independent Lasso paths with $\mathbb{P}(\Pi = 50) = 1 - \mathbb{P}(\Pi = 0) = \epsilon$. (b) Mean FDP vs. mean TPP averaged at different values of λ over 100 replicates for $n = p = 1000$, $\mathbb{P}(\Pi = 0) = 1 - \epsilon$ as before, and $\mathbb{P}(\Pi = 50 | \Pi \neq 0) = 1 - \mathbb{P}(\Pi = 0.1 | \Pi \neq 0) = \epsilon'$.

We think of M as being very large so that the (nonzero) signals are either very strong or very weak. In Appendix B, we prove that for some fixed $\epsilon' > 0$,

$$\lim_{M \to \infty} \lim_{n,p \to \infty} (\mathrm{TPP}(\lambda), \mathrm{FDP}(\lambda)) \to (u, q^\star(u)), \tag{2.5}$$

where convergence occurs in probability. This holds provided that $\lambda \to \infty$ in such a way that $M/\lambda \to \infty$; e.g. $\lambda = \sqrt{M}$. Hence, the most favorable configuration is when the signal is a mixture of very strong and very weak effect sizes because weak effects cannot be counted as false positives, thus reducing the FDP.

Figure 5b provides an illustration of (2.5). The setting is as in Figure 5a with $n = p = 1000$ and $\mathbb{P}(\Pi = 0) = 1 - \epsilon$ except that, here, conditionally on being nonzero the prior takes on the values 50 and 0.1 with probability $\epsilon' \in \{0.3, 0.5, 0.7, 0.9\}$ and $1 - \epsilon'$, respectively, so that we have a mixture of strong and weak signals. We observe that the true/false positive rate curve nicely touches *only* one point on the boundary depending on the proportion ϵ' of strong signals .

2.4 Technical novelties

The proof of Theorem 1 is built on top of the approximate message passing (AMP) theory developed in [9, 2, 1], and requires nontrivial extensions. AMP was originally designed as an algorithmic solution to compressive sensing problems under random Gaussian designs. In recent years, AMP has also found applications in robust statistics [10, 11], structured principal component analysis [8, 21], and the analysis of the stochastic block model [7]. Having said this, AMP theory is of crucial importance to us because it turns out to be a very useful technique to rigorously study various

statistical properties of the Lasso solution whenever we employ a *fixed* value of the regularizing parameter λ [3, 20, 22].

There are, however, major differences between our work and AMP research. First and foremost, our paper is concerned with practical research and conclusions that can be obtained when selecting λ adaptively, i.e. from the data; this is clearly outside of the envelope of current AMP results. Second, we are also concerned with situations where the noise variance can be zero. Likewise, this is outside of current knowledge. These differences are significant and as far as we know, our main result cannot be seen as a straightforward extension of AMP theory. In particular, we introduce a host of novel elements to deal, for instance, with the *irregularity* of the Lasso path. The irregularity means that a variable can enter and leave the model multiple times along the Lasso path [14, 29] so that natural sequences of Lasso models are not nested. This implies that a naive application of sandwiching inequalities does not give the type of statements holding uniformly over all λ's that we are looking for.

Instead, we develop new tools to understand the "continuity" of the support of $\widehat{\boldsymbol{\beta}}(\lambda)$ as a function of λ. Since the support can be characterized by the Karush-Kuhn-Tucker (KKT) conditions, this requires establishing some sort of continuity of the KKT conditions. Ultimately, we shall see that this comes down to understanding the maximum distance—uniformly over λ and λ'—between Lasso estimates $\widehat{\boldsymbol{\beta}}(\lambda)$ and $\widehat{\boldsymbol{\beta}}(\lambda')$ at close values of the regularizing parameter.

3 What's Wrong with Shrinkage?

We pause here to discuss the cause underlying the limitations of the Lasso for variable selection, which comes from the pseudo-noise introduced by shrinkage. As is well-known, the Lasso applies some form of soft-thresholding. This means that when the regularization parameter λ is large, the Lasso estimates are seriously biased downwards. Another way to put this is that the residuals still contain much of the effects associated with the selected variables. This can be thought of as extra noise that we may want to call *shrinkage noise*. Now as many strong variables get picked up, the shrinkage noise gets inflated and its projection along the directions of some of the null variables may actually dwarf the signals coming from the strong regression coefficients; this is why null variables get picked up. Although our exposition below lacks in rigor, it nevertheless formalizes this point in some *qualitative* fashion. It is important to note, however, that this phenomenon occurs in the linear sparsity regime considered in this paper so that we have sufficiently many variables for the shrinkage noise to build up and have a fold on other variables that become competitive with the signal. In contrast, under extreme sparsity and high SNR, both type I and II errors can be controlled at low levels, see e.g. [17].

For simplicity, we fix the true support $\mathcal{T}$ to be a deterministic subset of size $\epsilon \cdot p$, each nonzero coefficient in $\mathcal{T}$ taking on a constant value $M > 0$. Also, assume $\delta > \epsilon$. Finally, since the noiseless case $\boldsymbol{z} = \mathbf{0}$ is conceptually perhaps the most difficult, suppose $\sigma = 0$. Consider the reduced Lasso problem first:

$$\widehat{\boldsymbol{\beta}}_{\mathcal{T}}(\lambda) = \underset{\boldsymbol{b}_{\mathcal{T}} \in \mathbb{R}^{\epsilon p}}{\operatorname{argmin}} \ \tfrac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}_{\mathcal{T}}\boldsymbol{b}_{\mathcal{T}}\|^2 + \lambda \|\boldsymbol{b}_{\mathcal{T}}\|_1.$$

This (reduced) solution $\widehat{\boldsymbol{\beta}}_{\mathcal{T}}(\lambda)$ is independent from the other columns $\boldsymbol{X}_{\overline{\mathcal{T}}}$ (here and below $\overline{\mathcal{T}}$ is the complement of $\mathcal{T}$). Now take λ to be of the same magnitude as M so that roughly half of the signal variables are selected. The KKT conditions state that

$$-\lambda \mathbf{1} \le \boldsymbol{X}_{\mathcal{T}}^{\top}(\boldsymbol{y} - \boldsymbol{X}_{\mathcal{T}}\widehat{\boldsymbol{\beta}}_{\mathcal{T}}) \le \lambda \mathbf{1},$$

where $\mathbf{1}$ is the vectors of all ones. Note that if $|\boldsymbol{X}_j^\top(\boldsymbol{y} - \boldsymbol{X}_\mathcal{T}\widehat{\boldsymbol{\beta}}_\mathcal{T})| \le \lambda$ for all $j \in \overline{\mathcal{T}}$, then extending $\widehat{\boldsymbol{\beta}}_\mathcal{T}(\lambda)$ with zeros would be the solution to the full Lasso problem—with all variables included as potential predictors—since it would obey the KKT conditions for the full problem. A first simple fact is this: for $j \in \overline{\mathcal{T}}$, if

$$|\boldsymbol{X}_j^\top(\boldsymbol{y} - \boldsymbol{X}_\mathcal{T}\widehat{\boldsymbol{\beta}}_\mathcal{T})| > \lambda, \tag{3.1}$$

then $\boldsymbol{X}_j$ must be selected by the incremental Lasso with design variables indexed by $\mathcal{T} \cup \{j\}$. Now we make an assumption which is heuristically reasonable: any j obeying (3.1) has a reasonable chance to be selected in the full Lasso problem with the same λ (by this, we mean with some probability bounded away from zero). We argue in favor of this heuristic later.

Following our heuristic, we would need to argue that (3.1) holds for a number of variables in $\overline{\mathcal{T}}$ *linear* in p. Write

$$\boldsymbol{X}_\mathcal{T}^\top(\boldsymbol{y} - \boldsymbol{X}_\mathcal{T}\widehat{\boldsymbol{\beta}}_\mathcal{T}) = \boldsymbol{X}_\mathcal{T}^\top(\boldsymbol{X}_\mathcal{T}\boldsymbol{\beta}_\mathcal{T} - \boldsymbol{X}_\mathcal{T}\widehat{\boldsymbol{\beta}}_\mathcal{T}) = \lambda \boldsymbol{g}_\mathcal{T},$$

where $\boldsymbol{g}_\mathcal{T}$ is a subgradient of the ℓ_1 norm at $\widehat{\boldsymbol{\beta}}_\mathcal{T}$. Hence, $\boldsymbol{\beta}_\mathcal{T} - \widehat{\boldsymbol{\beta}}_\mathcal{T} = \lambda(\boldsymbol{X}_\mathcal{T}^\top\boldsymbol{X}_\mathcal{T})^{-1}\boldsymbol{g}_\mathcal{T}$ and

$$\boldsymbol{X}_\mathcal{T}(\boldsymbol{\beta}_\mathcal{T} - \widehat{\boldsymbol{\beta}}_\mathcal{T}) = \lambda \boldsymbol{X}_\mathcal{T}(\boldsymbol{X}_\mathcal{T}^\top\boldsymbol{X}_\mathcal{T})^{-1}\boldsymbol{g}_\mathcal{T}.$$

Since $\delta > \epsilon$, $\boldsymbol{X}_\mathcal{T}(\boldsymbol{X}_\mathcal{T}^\top\boldsymbol{X}_\mathcal{T})^{-1}$ has a smallest singular value bounded away from zero (since $\boldsymbol{X}_\mathcal{T}$ is a fixed random matrix with more rows than columns). Now because we make about half discoveries, the subgradient takes on the value one (in magnitude) at about $\epsilon \cdot p/2$ times. Hence, with high probability,

$$\|\boldsymbol{X}_\mathcal{T}(\boldsymbol{\beta}_\mathcal{T} - \widehat{\boldsymbol{\beta}}_\mathcal{T})\| \ge \lambda \cdot c_0 \cdot \|\boldsymbol{g}_\mathcal{T}\| \ge \lambda \cdot c_1 \cdot p$$

for some constants c_0, c_1 depending on ϵ and δ.

Now we use the fact that $\widehat{\boldsymbol{\beta}}_\mathcal{T}(\lambda)$ is independent of $\boldsymbol{X}_{\overline{\mathcal{T}}}$. For any $j \notin \mathcal{T}$, it follows that

$$\boldsymbol{X}_j^\top(\boldsymbol{y} - \boldsymbol{X}_\mathcal{T}\widehat{\boldsymbol{\beta}}_\mathcal{T}) = \boldsymbol{X}_j^\top\boldsymbol{X}_\mathcal{T}(\boldsymbol{\beta}_\mathcal{T} - \widehat{\boldsymbol{\beta}}_\mathcal{T})$$

is conditionally normally distributed with mean zero and variance

$$\frac{\|\boldsymbol{X}_\mathcal{T}(\boldsymbol{\beta}_\mathcal{T} - \widehat{\boldsymbol{\beta}}_\mathcal{T})\|^2}{n} \ge \frac{c_1\lambda^2 p}{n} = c_2 \cdot \lambda^2.$$

In conclusion, the probability that $\boldsymbol{X}_j^\top(\boldsymbol{y} - \boldsymbol{X}_\mathcal{T}\widehat{\boldsymbol{\beta}}_\mathcal{T})$ has absolute value larger than λ is bounded away from 0. Since there are $(1-\epsilon)p$ such j's, their expected number is linear in p. This implies that by the time half of the true variables are selected, we already have a non-vanishing FDP. Note that when $|\mathcal{T}|$ is not linear in p but smaller, e.g. $|\mathcal{T}| \le c_0 n/\log p$ for some sufficiently small constant c_0, the variance is much smaller because the estimation error $\|\boldsymbol{X}_\mathcal{T}(\boldsymbol{\beta}_\mathcal{T} - \widehat{\boldsymbol{\beta}}_\mathcal{T})\|^2$ is much lower, and this phenomenon does not occur.

Returning to our heuristic, we make things simpler by considering alternatives: (a) if very few extra variables in $\overline{\mathcal{T}}$ were selected by the full Lasso, then the value of the prediction $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$ would presumably be close to that obtained from the reduced model. In other words, the residuals $\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ from the full problem should not differ much from those in the reduced problem. Hence, for any j obeying (3.1), $\boldsymbol{X}_j$ would have a high correlation with $\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$. Thus this correlation has a good chance to be close to λ, or actually be equal to λ. Equivalently, $\boldsymbol{X}_j$ would likely be selected by the full Lasso problem. (b) If on the other hand, the number of variables selected from $\overline{\mathcal{T}}$ by the full Lasso were a sizeable proportion of $|\mathcal{T}|$, we would have lots of false discoveries, which is our claim.

In a more rigorous way, AMP claims that under our working hypothesis, the Lasso estimates $\widehat{\beta}_j(\lambda)$ are, in a certain sense, asymptotically distributed as $\eta_{\alpha\tau}(\beta_j + \tau W_j)$ for most j and W_j's independently drawn from $\mathcal{N}(0,1)$. The positive constants α and τ are uniquely determined by a pair of nonlinear equations parameterized by $\epsilon, \delta, \Pi, \sigma^2$, and λ. Suppose as before that all the nonzero coefficients of $\boldsymbol{\beta}$ are large in magnitude, say they are all equal to M. When about half of them appear on the path, we have that λ is just about equal to M. A consequence of the AMP equations is that τ is also of this order of magnitude. Hence, under the null we have that $(\beta_j + \tau W_j)/M \sim \mathcal{N}(0, (\tau/M)^2)$ while under the alternative, it is distributed as $\mathcal{N}(1, (\tau/M)^2)$. Because, τ/M is bounded away from zero, we see that false discoveries are bound to happen.

Variants of the Lasso and other ℓ_1-penalized methods, including ℓ_1-penalized logistic regression and the Dantzig selector, also suffer from this "shrinkage to noise" issue. However, this does not imply an ultimate limit of performance for all methods. If the signals are sufficiently strong, some other methods, perhaps with exponential computational cost, can achieve good model selection performance, see e.g. [24]. As an example, consider the simple ℓ_0-penalized maximum likelihood estimate,

$$\widehat{\boldsymbol{\beta}}_0 = \underset{\boldsymbol{b}\in\mathbb{R}^p}{\operatorname{argmin}} \ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2 + \lambda \|\boldsymbol{b}\|_0. \tag{3.2}$$

Methods known as AIC, BIC and RIC (short for risk inflation criterion) are all of this type and correspond to distinct values of λ. Then such fitting strategies can achieve perfect separation in some cases.

Theorem 2. *Under our working hypothesis, take $\epsilon < \delta$ for identifiability, and consider the two-point prior*

$$\Pi = \begin{cases} M, & w.p.\ \epsilon, \\ 0, & w.p.\ 1-\epsilon. \end{cases}$$

Then we can find $\lambda(M)$ such that in probability, the discoveries of the ℓ_0 estimator (3.2) *obey*

$$\lim_{M\to\infty} \lim_{n,p\to\infty} \mathrm{FDP} = 0 \quad \textit{and} \quad \lim_{M\to\infty} \lim_{n,p\to\infty} \mathrm{TPP} = 1.$$

Similar conclusions will certainly hold for many other non-convex methods, including SCAD and MC+ with properly tuned parameters [15, 34].

4 Discussion

We have evidenced a clear trade-off between false and true positive rates under the assumption that the design matrix has i.i.d. Gaussian entries. It is likely that there would be extensions of this result to designs with general i.i.d. sub-Gaussian entries as strong evidence suggests that the AMP theory may be valid for such larger classes, see [1]. It might also be of interest to study the Lasso trade-off diagram under correlated random designs.

Much literature on sparse regression has focused on the case where the number k of signals is at most a numerical constant times $n/\log p$ or $n/\log(p/k)$, showing support recovery under various assumptions on the design and the distribution of signal amplitudes. When n and p tend to infinity as in our setup, this means that k/p tends to zero, which is different from the linear sparsity regime discussed in this work. With a moderate degree of sparsity, the strong shrinkage noise makes the Lasso variable selection less accurate. This implies that model selection procedures derived by

stopping the Lasso path at some point (see e.g. [16, 26]) would lose power even for very strong signals. Thus it would be of interest to design and study other variable selection techniques to remedy this problem; e.g. by addressing the bias in Lasso estimates.

Of concern in this paper are statistical properties regarding the number of true and false discoveries along the Lasso path but it would also be interesting to study perhaps finer questions such as this: when does the first noise variable get selected? Consider Figure 6: there, the setting is the same as in Figure 1 before (high SNR regime) except that the number k of signals now varies from 5 to 150. In the very low sparsity regime, all the signal variables are selected before any noise variable. Once the sparsity passes a threshold, the first discovery keeps on occurring earlier as k continues to increase. The turning point is around $k = n/(2\log p)$, which is inline with some of the findings in [31] although this latter reference does not discuss the other aspect of this plot; namely the fact that the curve decreases after the turning point. In the linear sparsity regime, it would be interesting to derive a prediction for the time of first entry, in the case where the signals are large, say.

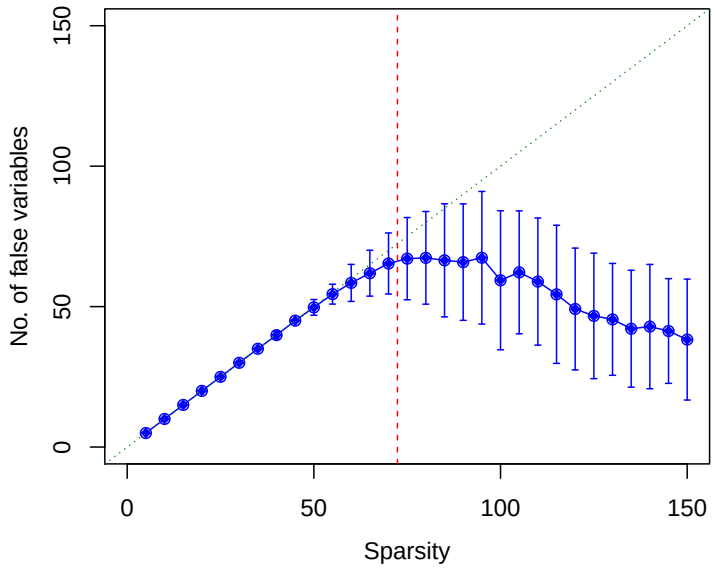


Figure 6: Rank of the first false discovery. Here, $n = p = 1000$ and $\beta_1 = \cdots = \beta_k = 50$ for k ranging from 5 to 150 ($\beta_i = 0$ for $i > k$). We plot averages from 100 independent replicates and display 1-SE error bars. The vertical line is placed at $k = n/(2\log p)$ and the 45° line passing through the origin is shown for convenience.

Acknowledgements

W. S. was partially supported by a General Wang Yaowu Stanford Graduate Fellowship. M. B. was partially supported by the statutory grants S30028/K1101 and S40067/K1101 at Wroclaw University of Technology. E. C. was partially supported by NSF under grant CCF-0963835, and by the Math + X Award from the Simons Foundation. W. S. would like to thank Andrea Montanari for helpful discussions. We thank Yuxin Chen for helpful comments about an early version of the manuscript.

References

[1] M. Bayati, M. Lelarge, and A. Montanari. Universality in polytope phase transitions and message passing algorithms. *The Annals of Applied Probability*, 25(2):753–822, 2015.

[2] M. Bayati and A. Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Trans. Inform. Theory*, 57(2):764–785, 2011.

[3] M. Bayati and A. Montanari. The Lasso risk for Gaussian matrices. *IEEE Trans. Inform. Theory*, 58(4):1997–2017, 2012.

[4] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009.

[5] M. Bogdan, E. v. d. Berg, W. Su, and E. J. Candès. Supplementary material to "Statistical estimation and testing via the ordered ℓ_1 norm". Available at `http://statweb.stanford.edu/~candes/papers/SortedL1_SM.pdf`, 2013.

[6] E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, Dec. 2005.

[7] Y. Deshpande, E. Abbe, and A. Montanari. Asymptotic mutual information for the two-groups stochastic block model. *arXiv preprint arXiv:1507.08685*, 2015.

[8] Y. Deshpande and A. Montanari. Information-theoretically optimal sparse PCA. In *IEEE International Symposium on Information Theory (ISIT)*, pages 2197–2201, 2014.

[9] D. L. Donoho, A. Maleki, and A. Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

[10] D. L. Donoho and A. Montanari. High dimensional robust M-estimation: Asymptotic variance via approximate message passing. *arXiv preprint arXiv:1310.7320*, 2013.

[11] D. L. Donoho and A. Montanari. Variance breakdown of Huber (M)-estimators: $n/p \to m \in (1, \infty)$. *arXiv preprint arXiv:1503.02106*, 2015.

[12] D. L. Donoho and J. Tanner. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Trans. R. Soc. A*, 367(1906):4273–4293, 2009.

[13] L. S. Eberlin et al. Molecular assessment of surgical-resection margins of gastric cancer by mass-spectrometric imaging. *Proceedings of the National Academy of Sciences*, 111(7):2436–2441, 2014.

[14] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

[15] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[16] M. G. G'Sell, S. Wager, A. Chouldechova, and R. Tibshirani. Sequential selection procedures and false discovery rate control. *arXiv preprint arXiv:1309.5352*, 2013.

[17] P. Ji and Z. Zhao. Rate optimal multiple testing procedure in high-dimensional regression. *arXiv preprint arXiv:1404.2961*, 2014.

[18] W. Lee et al. A high-resolution atlas of nucleosome occupancy in yeast. *Nature genetics*, 39(10):1235–1244, 2007.

[19] R. Lockhart, J. Taylor, R. J. Tibshirani, and R. Tibshirani. A significance test for the lasso. *The Annals of Statistics*, 42(2):413, 2014.

[20] A. Maleki, L. Anitori, Z. Yang, and R. Baraniuk. Asymptotic analysis of complex LASSO via complex approximate message passing (CAMP). *IEEE Trans. Inform. Theory*, 59(7):4290–4308, 2013.

[21] A. Montanari and E. Richard. Non-negative principal component analysis: Message passing algorithms and sharp asymptotics. *arXiv preprint arXiv:1406.4775*, 2014.

[22] A. Mousavi, A. Maleki, and R. G. Baraniuk. Asymptotic analysis of LASSO's solution path with implications for approximate message passing. *arXiv preprint arXiv:1309.5979*, 2013.

[23] F. S. Paolo, H. A. Fricker, and L. Padman. Volume loss from Antarctic ice shelves is accelerating. *Science*, 348(6232):327–331, 2015.

[24] G. Reeves and M. C. Gastpar. Approximate sparsity pattern recovery: Information-theoretic lower bounds. *IEEE Trans. Inform. Theory*, 59(6):3451–3465, 2013.

[25] F. Santosa and W. W. Symes. Linear inversion of band-limited reflection seismograms. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1307–1330, 1986.

[26] J. Taylor and R. Tibshirani. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634, 2015.

[27] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, Feb. 1994.

[28] R. J. Tibshirani. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490, 2013.

[29] R. J. Tibshirani and J. Taylor. The solution path of the generalized lasso. *The Annals of Statistics*, 39(3):1335–1371, 2011.

[30] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing, Theory and Applications*, 2012.

[31] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso). *IEEE Trans. Inform. Theory*, 55(5):2183–2202, 2009.

[32] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67, 2006.

[33] Y. Yuan et al. Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nature biotechnology*, 32(7):644–652, 2014.

[34] C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

[35] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

[36] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.

A For all Values of λ Simultaneousy

Throughout, we work under our working hypothesis and, for the moment, take $\sigma > 0$. Also, we shall frequently use results in [3], notably, Theorem 1.5, Lemma 3.1, Lemma 3.2, and Proposition 3.6 therein. Having said this, most of our proofs are rather self-contained, and the strategies accessible to readers who have not yet read [3].

At a high level, the proof of uniform convergence has two distinct aspects: we first characterize the Lasso solution at fixed values of λ, and then exhibit uniformity of these results over λ. The first step is accomplished largely by resorting to off-the-shelf AMP theory, whereas the second step is new and presents technical challenges.

As a start, our first result below accurately predicts the asymptotic limits of FDP and power at a fixed λ. This lemma is borrowed from [5], which follows from Theorem 1.5 in [3] in a natural way, albeit with some effort spent in resolving a continuity issue near the origin. Recall that $\eta_t(\cdot)$ is the soft-thresholding operator defined as $\eta_t(x) = \mathrm{sgn}(x)(|x| - t)_+$, and $\Pi^\star$ is the distribution of Π conditionally on being nonzero;

$$\Pi = \begin{cases} \Pi^\star, & \text{w.p. } \epsilon, \\ 0, & \text{w.p. } 1-\epsilon. \end{cases}$$

Denote by α_0 the unique root of $(1+t^2)\Phi(-t) - t\phi(t) = \delta/2$.

Lemma A.1 (Theorem 1 in [5]; see also Theorem 1.5 in [3]). *The Lasso solution with a fixed $\lambda > 0$ obeys*

$$\frac{V(\lambda)}{p} \xrightarrow{\mathbb{P}} 2(1-\epsilon)\Phi(-\alpha), \quad \frac{T(\lambda)}{p} \xrightarrow{\mathbb{P}} \mathbb{P}(|\Pi + \tau W| > \alpha\tau, \Pi \neq 0) = \epsilon\,\mathbb{P}(|\Pi^\star + \tau W| > \alpha\tau),$$

where W is $\mathcal{N}(0,1)$ independent of Π, and $\tau > 0, \alpha > \max\{\alpha_0, 0\}$ is the unique solution to

$$\begin{aligned} \tau^2 &= \sigma^2 + \frac{1}{\delta}\mathbb{E}\left(\eta_{\alpha\tau}(\Pi + \tau W) - \Pi\right)^2 \\ \lambda &= \left(1 - \frac{1}{\delta}\,\mathbb{P}(|\Pi + \tau W| > \alpha\tau)\right)\alpha\tau. \end{aligned} \tag{A.1}$$

We briefly discuss how Lemma A.1 follows from Theorem 1.5 in [3]. There, it is rigorously proven that the joint distribution of $(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}})$ is, in some sense, asymptotically the same as that of $(\boldsymbol{\beta}, \eta_{\alpha\tau}(\boldsymbol{\beta} + \tau\boldsymbol{W}))$, where $\boldsymbol{W}$ is a p-dimensional vector of i.i.d. standard normals independent of $\boldsymbol{\beta}$, and where the soft-thresholding operation acts in a componentwise fashion. Roughly speaking, the Lasso estimate $\widehat{\beta}_j$ looks like $\eta_{\alpha\tau}(\beta_j + \tau W_j)$, so that we are applying soft thresholding at level $\alpha\tau$ rather than λ and the noise level is τ rather than σ. With these results in place, we informally get

$$\begin{aligned} V(\lambda)/p = \#\{j : \widehat{\beta}_j \neq 0, \beta_j = 0\}/p &\approx \mathbb{P}(\eta_{\alpha\tau}(\Pi + \tau W) \neq 0, \Pi = 0) \\ &= (1-\epsilon)\ \mathbb{P}(|\tau W| > \alpha\tau) \\ &= 2(1-\epsilon)\,\Phi(-\alpha). \end{aligned}$$

Similarly, $T(\lambda)/p \approx \epsilon\ \mathbb{P}(|\Pi^\star + \tau W| > \alpha\tau)$. For details, we refer the reader to Theorem 1 in [5].

Our interest is to extend this convergence result uniformly over a range of λ. The proof of the next result is the subject of Section A.1. Note that both $\tau = \tau_\lambda$ and $\alpha = \alpha_\lambda$ depend on λ.

Lemma A.2. *For any fixed $0 < \lambda_{\min} < \lambda_{\max}$, the convergence in Lemma A.1 is uniform over $\lambda \in [\lambda_{\min}, \lambda_{\max}]$. That is, setting*

$$\mathrm{fd}^{\infty}(\lambda) := 2(1-\epsilon)\Phi(-\alpha), \quad \mathrm{td}^{\infty}(\lambda) := \epsilon\,\mathbb{P}(|\Pi^{\star} + \tau W| > \alpha\tau) \tag{A.2}$$

we have

$$\sup_{\lambda_{\min} \le \lambda \le \lambda_{\max}} \left| \frac{V(\lambda)}{p} - \mathrm{fd}^{\infty}(\lambda) \right| \xrightarrow{\mathbb{P}} 0,$$

and

$$\sup_{\lambda_{\min} \le \lambda \le \lambda_{\max}} \left| \frac{T(\lambda)}{p} - \mathrm{td}^{\infty}(\lambda) \right| \xrightarrow{\mathbb{P}} 0.$$

A.1 Proof of Lemma A.2

Lemma A.3. *For any $c > 0$, there exists a constant $r_c > 0$ such that for any arbitrary $r > r_c$,*

$$\sup_{\|\boldsymbol{u}\|=1} \#\left\{ 1 \le j \le p : |\boldsymbol{X}_j^{\top}\boldsymbol{u}| > \frac{r}{\sqrt{n}} \right\} \le cp$$

holds with probability tending to one.

A key ingredient in the proof of Lemma A.2 is, in a certain sense, the uniform continuity of the support of $\widehat{\boldsymbol{\beta}}(\lambda)$. This step is justified by the auxiliary lemma below which demonstrates that the Lasso estimates are uniformly continuous in ℓ_2 norm.

Lemma A.4. *Fixe $0 < \lambda_{\min} < \lambda_{\max}$. Then there is a constant c such for any $\lambda^- < \lambda^+$ in $[\lambda_{\min}, \lambda_{\max}]$,*

$$\sup_{\lambda^- \le \lambda \le \lambda^+} \left\| \widehat{\boldsymbol{\beta}}(\lambda) - \widehat{\boldsymbol{\beta}}(\lambda^-) \right\| \le c\sqrt{(\lambda^+ - \lambda^-)p}$$

holds with probability tending to one.

Proof of Lemma A.2. We prove the uniform convergence of $V(\lambda)/p$ and similar arguments apply to $T(\lambda)/p$. To begin with, let $\lambda_{\min} = \lambda_0 < \lambda_1 < \cdots < \lambda_m = \lambda_{\max}$ be equally spaced points and set $\Delta := \lambda_{i+1} - \lambda_i = (\lambda_{\max} - \lambda_{\min})/m$; the number of knots m shall be specified later. It follows from Lemma A.1 that

$$\max_{0 \le i \le m} |V(\lambda_i)/p - \mathrm{fd}^{\infty}(\lambda_i)| \xrightarrow{\mathbb{P}} 0 \tag{A.3}$$

by a union bound. For any constant $\omega > 0$, the continuity of $\mathrm{fd}^{\infty}(\lambda)$ as a function of λ on $[\lambda_{\min}, \lambda_{\max}]$ gives that

$$|\mathrm{fd}^{\infty}(\lambda) - \mathrm{fd}^{\infty}(\lambda')| \le \omega \tag{A.4}$$

holds for all $\lambda_{\min} \le \lambda, \lambda' \le \lambda_{\max}$ satisfying $|\lambda - \lambda'| \le 1/m$ provided m is sufficiently large. We now aim to show that if m is sufficiently large (but fixed), then

$$\max_{0 \le i < m} \sup_{\lambda_i \le \lambda \le \lambda_{i+1}} |V(\lambda)/p - V(\lambda_i)/p| \le \omega \tag{A.5}$$

holds with probability approaching one as $p \to \infty$. Since ω is arbitrary small, combining (A.3), (A.4), and (A.5) gives uniform convergence by applying the triangle inequality.

Let $\mathcal{S}(\lambda)$ be a short-hand for $\mathrm{supp}(\widehat{\boldsymbol{\beta}}(\lambda))$. Fix $0 \le i < m$ and put $\lambda^- = \lambda_i$ and $\lambda^+ = \lambda_{i+1}$. For any $\lambda \in [\lambda^-, \lambda^+]$,

$$|V(\lambda) - V(\lambda^-)| \le |\mathcal{S}(\lambda) \setminus \mathcal{S}(\lambda^-)| + |\mathcal{S}(\lambda^-) \setminus \mathcal{S}(\lambda)|. \tag{A.6}$$

Hence, it suffices to give upper bound about the sizes of $\mathcal{S}(\lambda) \setminus \mathcal{S}(\lambda^-)$ and $\mathcal{S}(\lambda^-) \setminus \mathcal{S}(\lambda)$. We start with $|\mathcal{S}(\lambda) \setminus \mathcal{S}(\lambda^-)|$.

The KKT optimality conditions for the Lasso solution state that there exists a subgradient $\boldsymbol{g}(\lambda) \in \partial \|\widehat{\boldsymbol{\beta}}(\lambda)\|_1$ obeying

$$\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda)) = \lambda \boldsymbol{g}(\lambda)$$

for each λ. Note that $g_j(\lambda) = \pm 1$ if $j \in \mathcal{S}(\lambda)$. As mentioned earlier, our strategy is to establish some sort of continuity of the KKT conditions with respect to λ. To this end, let

$$\boldsymbol{u} = \frac{\boldsymbol{X}\left(\widehat{\boldsymbol{\beta}}(\lambda) - \widehat{\boldsymbol{\beta}}(\lambda^-)\right)}{\left\|\boldsymbol{X}\left(\widehat{\boldsymbol{\beta}}(\lambda) - \widehat{\boldsymbol{\beta}}(\lambda^-)\right)\right\|}$$

be a point in $\mathbb{R}^n$ with unit ℓ_2 norm. Then for each $j \in \mathcal{S}(\lambda) \setminus \mathcal{S}(\lambda^-)$, we have

$$|\boldsymbol{X}_j^\top \boldsymbol{u}| = \frac{\left|\boldsymbol{X}_j^\top \boldsymbol{X}(\widehat{\boldsymbol{\beta}}(\lambda) - \widehat{\boldsymbol{\beta}}(\lambda^-))\right|}{\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}}(\lambda) - \widehat{\boldsymbol{\beta}}(\lambda^-))\|} = \frac{|\lambda g_j(\lambda) - \lambda^- g_j(\lambda^-)|}{\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}}(\lambda) - \widehat{\boldsymbol{\beta}}(\lambda^-))\|} \ge \frac{\lambda - \lambda^- |g_j(\lambda^-)|}{\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}}(\lambda) - \widehat{\boldsymbol{\beta}}(\lambda^-))\|}.$$

Now, given an arbitrary constant $a > 0$ to be determined later, either $|g_j(\lambda^-)| \in [1-a, 1)$ or $|g_j(\lambda^-)| \in [0, 1-a)$. In the first case ((a) below) note that we exclude $|g_j(\lambda^-)| = 1$ because for random designs, when $j \notin \mathcal{S}(\lambda)$ the equality $|\boldsymbol{X}_j^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda^-))| = \lambda^-$ can only hold with zero probability (see e.g. [28]). Hence, at least one of the following statements hold:

(a) $|\boldsymbol{X}_j^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda^-))| = \lambda^- |g_j(\lambda^-)| \in [(1-a)\lambda^-, \lambda^-)$;

(b) $|\boldsymbol{X}_j^\top \boldsymbol{u}| \ge \frac{\lambda - (1-a)\lambda^-}{\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}}(\lambda) - \widehat{\boldsymbol{\beta}}(\lambda^-))\|} > \frac{a\lambda^-}{\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}}(\lambda) - \widehat{\boldsymbol{\beta}}(\lambda^-))\|}$.

In the second case, since the spectral norm $\sigma_{\max}(\boldsymbol{X})$ is bounded in probability (see e.g. [30]), we make use of Lemma A.4 to conclude that

$$\frac{a\lambda^-}{\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}}(\lambda) - \widehat{\boldsymbol{\beta}}(\lambda^-))\|} \ge \frac{a\lambda^-}{\sigma_{\max}(\boldsymbol{X})\|\widehat{\boldsymbol{\beta}}(\lambda) - \widehat{\boldsymbol{\beta}}(\lambda^-)\|} \ge \frac{a\lambda^-}{c\sigma_{\max}(\boldsymbol{X})\sqrt{(\lambda^+ - \lambda^-)p}} \ge c' a \sqrt{\frac{m}{n}}$$

holds for all $\lambda^- \le \lambda \le \lambda^+$ with probability tending to one. Above, the constant c' only depends on $\lambda_{\min}, \lambda_{\max}, \delta$ and Π. Consequently, we see that

$$\sup_{\lambda^- \le \lambda \le \lambda^+} \left|\mathcal{S}(\lambda) \setminus \mathcal{S}(\lambda^-)\right| \le \#\left\{j : (1-a)\lambda^- \le |\boldsymbol{X}_j^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda^-))| < \lambda^-\right\} + \#\left\{j : |\boldsymbol{X}_j^\top \boldsymbol{u}| > c' a \sqrt{m/n}\right\}.$$

Equality (3.21) of [3] guarantees the existence of a constant a such that the event[3]

$$\#\left\{j : (1-a)\lambda^- \le |\boldsymbol{X}_j^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda^-))| < \lambda^-\right\} \le \frac{\omega p}{4} \tag{A.7}$$

[3]Apply Theorem 1.8 to carry over the results for AMP iterates to Lasso solution.

happens with probability approaching one. Since $\lambda^- = \lambda_i$ is always in the interval $[\lambda_{\min}, \lambda_{\max}]$, the constant a can be made to be independent of the index i. For the second term, it follows from Lemma A.3 that for sufficiently large m, the event

$$\#\left\{j : |\boldsymbol{X}_j^\top \boldsymbol{u}| > c' a\sqrt{m/n}\right\} \leq \frac{\omega p}{4} \tag{A.8}$$

also holds with probability approaching one. Combining (A.7) and (A.8), we get

$$\sup_{\lambda^- \leq \lambda \leq \lambda^+} \left|\mathcal{S}(\lambda) \setminus \mathcal{S}(\lambda^-)\right| \leq \frac{\omega p}{2} \tag{A.9}$$

holds with probability tending to one.

Next, we bound $|\mathcal{S}(\lambda^-) \setminus \mathcal{S}(\lambda)|$. Applying Theorem 1.5 in [3], we can find a constant $\nu > 0$ independent of $\lambda^- \in [\lambda_{\min}, \lambda_{\max}]$ such that

$$\#\left\{j : 0 < |\widehat{\beta}_j(\lambda^-)| < \nu\right\} \leq \frac{\omega p}{4} \tag{A.10}$$

happens with probability approaching one. Furthermore, the simple inequality

$$\|\widehat{\boldsymbol{\beta}}(\lambda) - \widehat{\boldsymbol{\beta}}(\lambda^-)\|^2 \geq \nu^2 \#\left\{j : j \in \mathcal{S}(\lambda^-) \setminus \mathcal{S}(\lambda), |\widehat{\beta}_j(\lambda^-)| \geq \nu\right\},$$

together with Lemma A.4, give

$$\#\left\{j : j \in \mathcal{S}(\lambda^-) \setminus \mathcal{S}(\lambda), |\widehat{\beta}_j(\lambda^-)| \geq \nu\right\} \leq \frac{\|\widehat{\boldsymbol{\beta}}(\lambda) - \widehat{\boldsymbol{\beta}}(\lambda^-)\|^2}{\nu^2} \leq \frac{c^2(\lambda^+ - \lambda^-)p}{\nu^2} \tag{A.11}$$

for all $\lambda \in [\lambda_{\min}, \lambda_{\max}]$ with probability converging to one. Taking m sufficiently large such that $\lambda^+ - \lambda^- = (\lambda_{\max} - \lambda_{\min})/m \leq \omega\nu^2/4c^2$ in (A.11) and combining this with (A.10) gives that

$$\sup_{\lambda^- \leq \lambda \leq \lambda^+} \left|\mathcal{S}(\lambda^-) \setminus \mathcal{S}(\lambda)\right| \leq \frac{\omega p}{2} \tag{A.12}$$

holds with probability tending to one.

To conclude the proof, note that both (A.9) and (A.12) hold for a large but fixed m. Substituting these two inequalities into (A.6) confirms (A.5) by taking a union bound.

As far as the true discovery number $T(\lambda)$ is concerned, all the arguments seamlessly apply and we do not repeat them. This terminates the proof. □

A.2 Proofs of auxiliary lemmas

In this section, we prove Lemmas A.3 and A.4. While the proof of the first is straightforward, the second crucially relies on Lemma A.7, whose proof makes use of Lemmas A.5 and A.6. Hereafter, we denote by $o_{\mathbb{P}}(1)$ any random variable which tends to zero in probability.

Proof of Lemma A.3. Since

$$\|\boldsymbol{u}^\top \boldsymbol{X}\|^2 \geq \frac{r^2}{n} \#\left\{1 \leq j \leq p : |\boldsymbol{X}_j^\top \boldsymbol{u}| > \frac{r}{\sqrt{n}}\right\},$$

we have

$$\# \left\{ 1 \le j \le p : |\boldsymbol{X}_j^\top \boldsymbol{u}| > \frac{r}{\sqrt{n}} \right\} \le \frac{n}{r^2} \|\boldsymbol{u}^\top \boldsymbol{X}\|^2 \le \frac{n\sigma_{\max}(\boldsymbol{X})^2 \|\boldsymbol{u}\|^2}{r^2}$$
$$= (1 + o_{\mathbb{P}}(1)) \frac{(1+\sqrt{\delta})^2 p}{r^2},$$

where we make use of $\lim n/p = \delta$ and $\sigma_{\max}(\boldsymbol{X}) = 1 + \delta^{-1/2} + o_{\mathbb{P}}(1)$. To complete the proof, take any $r_c > 0$ such that $(1+\sqrt{\delta})/r_c < \sqrt{c}$.

□

Lemma A.5. *Take a sequence $a_1 \ge a_2 \ge \cdots \ge a_p \ge 0$ with at least one strict inequality, and suppose that*

$$\frac{p \sum_{i=1}^p a_i^2}{\left(\sum_{i=1}^p a_i\right)^2} \ge M$$

for some $M > 1$. Then for any $1 \le s \le p$,

$$\frac{\sum_{i=1}^s a_i^2}{\sum_{i=1}^p a_i^2} \ge 1 - \frac{p^3}{Ms^3}.$$

Proof of Lemma A.5. By the monotonicity of $\boldsymbol{a}$,

$$\frac{\sum_{i=1}^s a_i^2}{s} \ge \frac{\sum_{i=1}^p a_i^2}{p},$$

which implies

$$\frac{p \sum_{i=1}^s a_i^2}{\left(\sum_{i=1}^s a_i\right)^2} \ge \frac{s \sum_{i=1}^p a_i^2}{\left(\sum_{i=1}^p a_i\right)^2} \ge \frac{sM}{p}. \tag{A.13}$$

Similarly,

$$\sum_{i=s+1}^p a_i^2 \le (p-s) \left(\frac{\sum_{i=1}^s a_i}{s} \right)^2 \tag{A.14}$$

and it follows from (A.13) and (A.14) that

$$\frac{\sum_{i=1}^s a_i^2}{\sum_{i=1}^p a_i^2} \ge \frac{\sum_{i=1}^s a_i^2}{\sum_{i=1}^s a_i^2 + (p-s)\left(\frac{\sum_{i=1}^s a_i}{s}\right)^2} \ge \frac{\frac{sM}{p^2}}{\frac{sM}{p^2} + \frac{p-s}{s^2}} \ge 1 - \frac{p^3}{Ms^3}.$$

□

Lemma A.6. *Assume $n/p \to 1$, i.e. $\delta = 1$. Suppose s obeys $s/p \to 0.01$. Then with probability tending to one, the smallest singular value obeys*

$$\min_{|\mathcal{S}|=s} \sigma_{\min}(\boldsymbol{X}_{\mathcal{S}}) \ge \frac{1}{2},$$

where the minimization is over all subsets of $\{1, \ldots, p\}$ of cardinality s.

Proof of Lemma A.6. For a fixed $\mathcal{S}$, we have

$$\mathbb{P}\left(\sigma_{\min}(\boldsymbol{X}_{\mathcal{S}}) < 1 - \sqrt{s/n} - t\right) \le \mathrm{e}^{-nt^2/2}$$

for all $t \ge 0$, please see [30]. The claim follows from plugging $t = 0.399$ and a union bound over $\binom{p}{s} \le \exp(pH(s/p))$ subsets, where $H(q) = -q\log q - (1-q)\log(1-q)$. We omit the details. □

The next lemma bounds the Lasso solution in ℓ_2 norm uniformly over λ. We use some ideas from the proof of Lemma 3.2 in [3].

Lemma A.7. *Given any positive constants $\lambda_{\min} < \lambda_{\max}$, there exists a constant C such that*

$$\mathbb{P}\left(\sup_{\lambda_{\min} \le \lambda \le \lambda_{\max}} \|\widehat{\boldsymbol{\beta}}(\lambda)\| \le C\sqrt{p}\right) \to 1.$$

Proof of Lemma A.7. For simplicity, we omit the dependency of $\widehat{\boldsymbol{\beta}}$ on λ when clear from context. We first consider the case where $\delta < 1$. Write $\widehat{\boldsymbol{\beta}} = \mathcal{P}_1(\widehat{\boldsymbol{\beta}}) + \mathcal{P}_2(\widehat{\boldsymbol{\beta}})$, where $\mathcal{P}_1(\widehat{\boldsymbol{\beta}})$ is the projection of $\widehat{\boldsymbol{\beta}}$ onto the null space of $\boldsymbol{X}$ and $\mathcal{P}_2(\widehat{\boldsymbol{\beta}})$ is the projection of $\widehat{\boldsymbol{\beta}}$ onto the row space of $\boldsymbol{X}$. By the rotational invariance of i.i.d. Gaussian vectors, the null space of $\boldsymbol{X}$ is a random subspace of dimension $p - n = (1 - \delta + o(1))p$ with uniform orientation. Since $\mathcal{P}_1(\widehat{\boldsymbol{\beta}})$ belongs to the null space, Kashin's Theorem (see Theorem F.1 in [3]) gives that with probability at least $1 - 2^{-p}$,

$$\begin{aligned}
\|\widehat{\boldsymbol{\beta}}\|^2 &= \|\mathcal{P}_1(\widehat{\boldsymbol{\beta}})\|^2 + \|\mathcal{P}_2(\widehat{\boldsymbol{\beta}})\|^2 \\
&\le c_1 \frac{\|\mathcal{P}_1(\widehat{\boldsymbol{\beta}})\|_1^2}{p} + \|\mathcal{P}_2(\widehat{\boldsymbol{\beta}})\|^2 \\
&\le 2c_1 \frac{\|\widehat{\boldsymbol{\beta}}\|_1^2 + \|\mathcal{P}_2(\widehat{\boldsymbol{\beta}})\|_1^2}{p} + \|\mathcal{P}_2(\widehat{\boldsymbol{\beta}})\|^2 \\
&\le \frac{2c_1\|\widehat{\boldsymbol{\beta}}\|_1^2}{p} + (1 + 2c_1)\|\mathcal{P}_2(\widehat{\boldsymbol{\beta}})\|^2
\end{aligned} \tag{A.15}$$

for some constant c_1 depending only on δ; the first step uses Kashin's theorem, the second the triangle inequality, and the third Cauchy-Schwarz inequality. The smallest nonzero singular value of the Wishart matrix $\boldsymbol{X}^\top\boldsymbol{X}$ is concentrated at $(1/\sqrt{\delta} - 1)^2$ with probability tending to one (see e.g. [30]). In addition, since $\mathcal{P}_2(\widehat{\boldsymbol{\beta}})$ belongs to the row space of $\boldsymbol{X}$, we have

$$\|\mathcal{P}_2(\widehat{\boldsymbol{\beta}})\|^2 \le c_2\|\boldsymbol{X}\mathcal{P}_2(\widehat{\boldsymbol{\beta}})\|^2$$

with probability approaching one. Above, c_2 can be chosen to be $(1/\sqrt{\delta} - 1)^{-2} + o(1)$. Set

$c_3 = c_2(1+2c_1)$. Continuing (A.15) yields

$$\begin{aligned}
\|\widehat{\boldsymbol{\beta}}\|^2 &\le \frac{2c_1\|\widehat{\boldsymbol{\beta}}\|_1^2}{p} + c_2(1+2c_1)\|\boldsymbol{X}\mathcal{P}_2(\widehat{\boldsymbol{\beta}})\|^2 \\
&= \frac{2c_1\|\widehat{\boldsymbol{\beta}}\|_1^2}{p} + c_3\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 \\
&\le \frac{2c_1\|\widehat{\boldsymbol{\beta}}\|_1^2}{p} + 2c_3\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 + 2c_3\|\boldsymbol{y}\|^2 \\
&\le \frac{2c_1\left(\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 + \lambda\|\widehat{\boldsymbol{\beta}}\|_1\right)^2}{\lambda^2 p} + 4c_3\left(\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 + \lambda\|\widehat{\boldsymbol{\beta}}\|_1\right) + 2c_3\|\boldsymbol{y}\|^2 \\
&\le \frac{c_1\|\boldsymbol{y}\|^4}{2\lambda^2 p} + 4c_3\|\boldsymbol{y}\|^2,
\end{aligned}$$

where in the last inequality we use the fact $\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2+\lambda\|\widehat{\boldsymbol{\beta}}\|_1 \le \frac{1}{2}\|\boldsymbol{y}\|^2$. Thus, it suffices to bound $\|\boldsymbol{y}\|^2$. The largest singular value of $\boldsymbol{X}^\top\boldsymbol{X}$ is bounded above by $(1/\sqrt{\delta}+1)^2 + o_{\mathbb{P}}(1)$. Therefore,

$$\|\boldsymbol{y}\|^2 = \|\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{z}\|^2 \le 2\|\boldsymbol{X}\boldsymbol{\beta}\|^2 + 2\|\boldsymbol{z}\|^2 \le c_4\|\boldsymbol{\beta}\|^2 + 2\|\boldsymbol{z}\|^2.$$

Since both β_i and z_i have bounded second moments, the law of large numbers claims that there exists a constant c_5 such that $c_4\|\boldsymbol{\beta}\|^2 + 2\|\boldsymbol{z}\|^2 \le c_5 p$ with probability approaching one. Combining all the inequalities above gives

$$\sup_{\lambda_{\min}\le\lambda\le\lambda_{\max}} \|\widehat{\boldsymbol{\beta}}(\lambda)\|^2 \le \frac{c_1 c_5^2 p}{2\lambda^2} + 2c_3c_5p \le \left(\frac{c_1c_5^2}{2\lambda_{\min}^2} + 2c_3c_5\right)p$$

with probability converging to one.

In the case where $\delta > 1$, the null space of $\boldsymbol{X}$ reduces to $\boldsymbol{0}$, hence $\mathcal{P}_1(\widehat{\boldsymbol{\beta}}) = 0$. Therefore, this reduces to a special case of the above argument.

Now, we turn to work on the case where $\delta = 1$. We start with

$$\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 \le 2\|\boldsymbol{y}\|^2 + 2\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2$$

and

$$\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 + \lambda\|\widehat{\boldsymbol{\beta}}\|_1 \le \frac{1}{2}\|\boldsymbol{y}\|^2.$$

These two inequalities give that simultaneously over all λ,

$$\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda)\|^2 \le 4\|\boldsymbol{y}\|^2 \le 4c_5p \tag{A.16a}$$

$$\|\widehat{\boldsymbol{\beta}}(\lambda)\|_1 \le \frac{1}{2\lambda_{\min}}\|\boldsymbol{y}\|^2 \le \frac{c_5 p}{2\lambda_{\min}} \tag{A.16b}$$

with probability converging to one. Let M be any constant larger than 1.7×10^7. If $p\|\widehat{\boldsymbol{\beta}}\|^2/\|\widehat{\boldsymbol{\beta}}\|_1^2 < M$, by (A.16b), we get

$$\|\widehat{\boldsymbol{\beta}}\| \le \frac{\sqrt{M}c_5}{2\lambda_{\min}}\sqrt{p}. \tag{A.17}$$

Otherwise, denoting by $\mathcal{T}$ the set of indices $1 \le i \le p$ that correspond to the $s := \lceil p/100 \rceil$ largest $|\widehat{\beta}|_i$, from Lemma A.5 we have

$$\frac{\|\widehat{\boldsymbol{\beta}}_{\mathcal{T}}\|^2}{\|\widehat{\boldsymbol{\beta}}\|^2} \ge 1 - \frac{p^3}{Ms^3} \ge 1 - \frac{10^6}{M}. \tag{A.18}$$

To proceed, note that

$$\begin{aligned}\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}\| &= \|\boldsymbol{X}_{\mathcal{T}}\widehat{\boldsymbol{\beta}}_{\mathcal{T}} + \boldsymbol{X}_{\overline{\mathcal{T}}}\widehat{\boldsymbol{\beta}}_{\overline{\mathcal{T}}}\| \\ &\ge \|\boldsymbol{X}_{\mathcal{T}}\widehat{\boldsymbol{\beta}}_{\mathcal{T}}\| - \|\boldsymbol{X}_{\overline{\mathcal{T}}}\widehat{\boldsymbol{\beta}}_{\overline{\mathcal{T}}}\| \\ &\ge \|\boldsymbol{X}_{\mathcal{T}}\widehat{\boldsymbol{\beta}}_{\mathcal{T}}\| - \sigma_{\max}(\boldsymbol{X})\|\widehat{\boldsymbol{\beta}}_{\overline{\mathcal{T}}}\|.\end{aligned}$$

By Lemma A.6, we get $\|\boldsymbol{X}_{\mathcal{T}}\widehat{\boldsymbol{\beta}}_{\mathcal{T}}\| \ge \frac{1}{2}\|\boldsymbol{\beta}_{\mathcal{T}}\|$, and it is also clear that $\sigma_{\max}(\boldsymbol{X}) = 2 + o_{\mathbb{P}}(1)$. Thus, by (A.18) we obtain

$$\begin{aligned}\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}\| \ge \|\boldsymbol{X}_{\mathcal{T}}\widehat{\boldsymbol{\beta}}_{\mathcal{T}}\| - \sigma_{\max}(\boldsymbol{X})\|\widehat{\boldsymbol{\beta}}_{\overline{\mathcal{T}}}\| &\ge \frac{1}{2}\|\widehat{\boldsymbol{\beta}}_{\mathcal{T}}\| - (2 + o_{\mathbb{P}}(1))\|\widehat{\boldsymbol{\beta}}_{\overline{\mathcal{T}}}\| \\ &\ge \frac{1}{2}\sqrt{1 - \frac{10^6}{M}}\|\widehat{\boldsymbol{\beta}}\| - (2 + o_{\mathbb{P}}(1))\sqrt{\frac{10^6}{M}}\|\widehat{\boldsymbol{\beta}}\| \\ &= (c + o_{\mathbb{P}}(1))\|\widehat{\boldsymbol{\beta}}\|,\end{aligned}$$

where $c = \frac{1}{2}\sqrt{1 - 10^6/M} - 2\sqrt{10^6/M} > 0$. Hence, owing to (A.16a),

$$\|\widehat{\boldsymbol{\beta}}\| \le \frac{(2 + o_{\mathbb{P}}(1))\sqrt{c_5}}{c}\sqrt{p} \tag{A.19}$$

In summary, with probability tending to one, in either case, namely, (A.17) or (A.19),

$$\|\widehat{\boldsymbol{\beta}}\| \le C\sqrt{p}$$

for come constant C. This holds uniformly for all $\lambda \in [\lambda_{\min}, \lambda_{\max}]$, and thus completes the proof. □

Now, we conclude this section by proving our main lemma.

Proof of Lemma A.4. The proof extensively applies Lemma 3.1[4] in [3] and Lemma A.7. Let $\boldsymbol{x}+\boldsymbol{r} = \widehat{\boldsymbol{\beta}}(\lambda)$ and $\boldsymbol{x} = \widehat{\boldsymbol{\beta}}(\lambda^-)$ be the notations in the statement of Lemma 3.1 in [3]. Among the fives assumptions needed in that lemma, it suffices to verify the first, third and fourth. Lemma A.7 asserts that

$$\sup_{\lambda^- \le \lambda \le \lambda^+} \|\boldsymbol{r}(\lambda)\| = \sup_{\lambda^- \le \lambda \le \lambda^+} \|\widehat{\boldsymbol{\beta}}(\lambda) - \widehat{\boldsymbol{\beta}}(\lambda^-)\| \le 2A\sqrt{p}$$

with probability approaching one. This fulfills the first assumption by taking $c_1 = 2A$. Next, let $\boldsymbol{g}(\lambda^-) \in \partial\|\widehat{\boldsymbol{\beta}}(\lambda^-)\|_1$ obey

$$\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda^-)) = \lambda^-\boldsymbol{g}(\lambda^-).$$

Hence,

$$\left\|-\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda^-)) + \lambda\boldsymbol{g}(\lambda^-)\right\| = (\lambda - \lambda^-)\|\boldsymbol{g}(\lambda^-)\| \le (\lambda^+ - \lambda^-)\sqrt{p}$$

[4]The conclusion of this lemma, $\|\boldsymbol{r}\| \le \sqrt{p}\xi(\epsilon, c_1, \ldots, c_5)$ in our notation, can be effortlessly strengthened to $\|\boldsymbol{r}\| \le (\sqrt{\epsilon} + \epsilon/\lambda)\xi(c_1, \ldots, c_5)\sqrt{p}$.

which certifies the third assumption. To verify the fourth assumption, taking $t \to \infty$ in Proposition 3.6 of [3] ensures the existence of constants c_2, c_3 and c_4 such that, with probability tending to one, $\sigma_{\min}(\boldsymbol{X}_{T\cup T'}) \geq c_4$ for $T = \{j : |g_j(\lambda^-)| \geq 1 - c_2\}$ and arbitrary $T' \subset \{1, \ldots, p\}$ with $|T'| \leq c_3 p$. Further, these constants can be independent of λ^- since $\lambda^- \in [\lambda_{\min}, \lambda_{\max}]$ belongs to a compact interval. Therefore, this lemma concludes that, with probability approaching one,

$$\begin{aligned}\sup_{\lambda^- \leq \lambda \leq \lambda^+} \|\widehat{\boldsymbol{\beta}}(\lambda) - \widehat{\boldsymbol{\beta}}(\lambda^-)\| &\leq \sup_{\lambda^- \leq \lambda \leq \lambda^+} \left(\sqrt{\lambda - \lambda^-} + \frac{\lambda - \lambda^-}{\lambda}\right) \xi\sqrt{p} \\ &\leq \left(\sqrt{\lambda^+ - \lambda^-} + \frac{\lambda^+ - \lambda^-}{\lambda_{\min}}\right) \xi\sqrt{p} \\ &= O\left(\sqrt{(\lambda^+ - \lambda^-)p}\right).\end{aligned}$$

This finishes the proof. □

B Optimizing the Tradeoff

In this section, we still work under our working hypothesis and $\sigma > 0$. Fixing δ and ϵ, we aim to show that all the pairs above the boundary curve can be realized. In the setting from Lemma A.1, define

$$\begin{aligned}\mathrm{tpp}^\infty(\lambda) &:= \mathbb{P}(|\Pi^\star + \tau W| > \alpha\tau) = \frac{\mathrm{td}^\infty(\lambda)}{\epsilon} \\ \mathrm{fdp}^\infty(\lambda) &:= \frac{2(1-\epsilon)\Phi(-\alpha)}{2(1-\epsilon)\Phi(-\alpha) + \epsilon\,\mathbb{P}(|\Pi^\star + \tau W| > \alpha\tau)} = \frac{\mathrm{fd}^\infty(\lambda)}{\mathrm{fd}^\infty(\lambda) + \mathrm{td}^\infty(\lambda)},\end{aligned}$$

to be the predicted FDP and TPP. (We shall often hide the dependence on λ.) Note that both tpp^∞ and fdp^∞ also depend on Π, δ, ϵ and σ.

Owing to the uniform convergence established in Appendix A, it is sufficient to study the range of $(\mathrm{tpp}^\infty(\lambda), \mathrm{fdp}^\infty(\lambda))$ by varying $\Pi^\star$ and λ. To this end, we introduce a useful trick based on the following lemma.

Lemma B.1. *For any fixed $\alpha > 0$, define a function $y = f(x)$ in the parametric form:*

$$\begin{aligned}x(t) &= \mathbb{P}(|t + W| > \alpha) \\ y(t) &= \mathbb{E}\,(\eta_\alpha(t + W) - t)^2\end{aligned}$$

for $t \geq 0$, where W is a standard normal. Then f is strictly concave.

We use this to simplify the problem of detecting feasible pairs $(\mathrm{tpp}^\infty, \mathrm{fdp}^\infty)$. Denote by $\pi^\star := \Pi^\star/\tau$. Then (A.1) implies

$$\begin{aligned}(1-\epsilon)\,\mathbb{E}\,\eta_\alpha(W)^2 + \epsilon\,\mathbb{E}(\eta_\alpha(\pi^\star + W) - \pi^\star)^2 &< \delta, \\ (1-\epsilon)\,\mathbb{P}(|W| > \alpha) + \epsilon\,\mathbb{P}(|\pi^\star + W| > \alpha) &< \min\{\delta, 1\}.\end{aligned} \tag{B.1}$$

We emphasize that (B.1) is not only necessary, but also sufficient in the following sense: given $0 < \delta_1 < \delta, 0 < \delta_2 < \min\{\delta, 1\}$ and $\pi^\star$, we can solve for τ by setting

$$(1-\epsilon)\,\mathbb{E}\,\eta_\alpha(W)^2 + \epsilon\,\mathbb{E}(\eta_\alpha(\pi^\star + W) - \pi^\star)^2 = \delta_1$$

and making use of the first line of (A.1), which can be alternatively written as

$$\tau^2 = \sigma^2 + \frac{\tau^2}{\delta}\left[(1-\epsilon)\,\mathbb{E}\,\eta_\alpha(W)^2 + \epsilon\,\mathbb{E}(\eta_\alpha(\pi^\star + W) - \pi^\star)^2\right].$$

($\Pi^\star = \tau\pi^\star$ is also determined, so does Π),[5] and along with

$$(1-\epsilon)\,\mathbb{P}(|W| > \alpha) + \epsilon\,\mathbb{P}(|\pi^\star + W| > \alpha) = \delta_2,$$

λ is also uniquely determined.

Since (B.1) is invariant if $\pi^\star$ is replaced by $|\pi^\star|$, we assume $\pi^\star \geq 0$ without loss of generality. As a function of t, $\mathbb{P}(|t+W| > \alpha)$ attains the minimum $\mathbb{P}(|W| > \alpha) = 2\Phi(-\alpha)$ at $t = 0$, and the supremum equal to one at $t = \infty$. Hence, there must exist $\epsilon' \in (0,1)$ obeying

$$\mathbb{P}(|\pi^\star + W| > \alpha) = (1-\epsilon')\,\mathbb{P}(|W| > \alpha) + \epsilon'. \tag{B.2}$$

As a consequence, the predicted TPP and FDP can be alternatively expressed as

$$\mathrm{tpp}^\infty = 2(1-\epsilon')\Phi(-\alpha) + \epsilon', \quad \mathrm{fdp}^\infty = \frac{2(1-\epsilon)\Phi(-\alpha)}{2(1-\epsilon\epsilon')\Phi(-\alpha) + \epsilon\epsilon'}. \tag{B.3}$$

Compared to the original formulation, this expression is preferred since it only involves scalars.

Now, we seek equations that govern ϵ' and α, given δ and ϵ. Since both $\mathbb{E}(\eta_\alpha(t+W) - t)^2$ and $\mathbb{P}(|t+W| > \alpha)$ are monotonically increasing with respect to $t \geq 0$, there exists a function f obeying

$$\mathbb{E}(\eta_\alpha(t+W) - t)^2 = f\left(\mathbb{P}(|t+W| > \alpha)\right).$$

Lemma B.1 states that f is concave. Then

$$\mathbb{E}(\eta_\alpha(\pi^\star + W) - \pi^\star)^2 = f\left(\mathbb{P}(|\pi^\star + W| > \alpha)\right.$$

and (B.2) allows us to view the argument of f in the right-hand side as an average of a random variable taking value $\mathbb{P}(|W| > \alpha)$ with probability $1-\epsilon'$ and value one with probability ϵ'. Therefore, Jensen's inequality states that

$$\mathbb{E}(\eta_\alpha(\pi^\star + W) - \pi^\star)^2 \geq (1-\epsilon')f\left(\mathbb{P}(|W| > \alpha)\right) + \epsilon' f(1) = (1-\epsilon')\,\mathbb{E}\,\eta_\alpha(W)^2 + \epsilon'(\alpha^2+1).$$

Combining this with (B.1) gives

$$(1-\epsilon\epsilon')\,\mathbb{E}\,\eta_\alpha(W)^2 + \epsilon\epsilon'(\alpha^2+1) < \delta, \tag{B.4a}$$

$$(1-\epsilon\epsilon')\,\mathbb{P}(|W| > \alpha) + \epsilon\epsilon' < \min\{\delta, 1\}. \tag{B.4b}$$

Similar to (B.1), (B.4) is also sufficient in the same sense, and (B.4b) is automatically satisfied if $\delta > 1$.

The remaining part of this section studies the range of $(\mathrm{tpp}^\infty, \mathrm{fdp}^\infty)$ given by (B.3) under the constraints (B.4). Before delving into the details, we remark that this reduction of $\pi^\star$ to a two-point prior is realized by setting $\pi^\star = \infty$ (equivalently $\Pi^\star = \infty$) with probability ϵ' and otherwise $+0$, where $+0$ is considered to be nonzero. Though this prior is not valid since the working hypothesis requires a finite second moment, it can nevertheless be approximated by a sequence of instances, please see the example given in Section 2.

The lemma below recognizes that for certain (δ, ϵ) pairs, the TPP is asymptotically bounded above away from 1.

[5] Not every pair $(\delta_1, \delta_2) \in (0,\delta)\times(0, \min\{\delta, 1\})$ is feasible below the Donoho-Tanner phase transition. Nevertheless, this does not affect our discussion.

Lemma B.2. *Put*

$$u^\star(\delta,\epsilon) := \begin{cases} 1 - \frac{(1-\delta)(\epsilon-\epsilon^\star)}{\epsilon(1-\epsilon^\star)}, & \delta < 1 \text{ and } \epsilon > \epsilon^\star(\delta), \\ 1, & \text{otherwise.} \end{cases}$$

Then

$$\mathrm{tpp}^\infty < u^\star(\delta,\epsilon).$$

Moreover, tpp^∞ *can be arbitrarily close to* $u^\star$*.*

This lemma directly implies that above the Donoho-Tanner phase transition (i.e. $\delta < 1$ and $\epsilon > \epsilon^\star(\delta)$), there is a fundamental limit on the TPP for arbitrarily strong signals. Consider

$$2(1-\epsilon)\left[(1+t^2)\Phi(-t) - t\phi(t)\right] + \epsilon(1+t^2) = \delta. \tag{B.5}$$

For $\delta < 1$, $\epsilon^\star$ is the only positive constant in $(0,1)$ such that (B.5) with $\epsilon = \epsilon^\star$ has a unique positive root. Alternatively, the function $\epsilon^\star = \epsilon^\star(\delta)$ is implicitly given in the following parametric form:

$$\begin{aligned} \delta &= \frac{2\phi(t)}{2\phi(t) + t(2\Phi(t)-1)} \\ \epsilon^\star &= \frac{2\phi(t) - 2t\Phi(-t)}{2\phi(t) + t(2\Phi(t)-1)} \end{aligned}$$

for $t > 0$, from which we see that $\epsilon^\star < \delta < 1$. Take the sparsity level k such as $\epsilon^\star p < k < \delta p = n$, from which we have $u^\star < 1$. As a result, the Lasso is unable to select all the k true signals even when the signal strength is arbitrarily high. This is happening even though the Lasso has the chance to select up to $n > k$ variables.

Any u between 0 and $u^\star$ (non-inclusive) can be realized as tpp^∞. Recall that we denote by $t^\star(u)$ the unique root in (α_0, ∞) (α_0 is the root of $(1+t^2)\Phi(-t) - t\phi(t) = \delta/2$) to the equation

$$\frac{2(1-\epsilon)\left[(1+t^2)\Phi(-t) - t\phi(t)\right] + \epsilon(1+t^2) - \delta}{\epsilon\left[(1+t^2)(1-2\Phi(-t)) + 2t\phi(t)\right]} = \frac{1-u}{1-2\Phi(-t)}. \tag{B.6}$$

For a proof of this fact, we refer to Lemma B.4. Last, recall that

$$q^\star(u;\delta,\epsilon) = \frac{2(1-\epsilon)\Phi(-t^\star(u))}{2(1-\epsilon)\Phi(-t^\star(u)) + \epsilon u}.$$

We can now state the fundamental tradeoff between fdp^∞ and tpp^∞.

Lemma B.3. *If* $\mathrm{tpp}^\infty \geq u$ *for* $u \in (0, u^\star)$*, then*

$$\mathrm{fdp}^\infty > q^\star(u).$$

In addition, fdp^∞ *can be arbitrarily close to* $q^\star(u)$*.*

B.1 Proofs of Lemmas B.1, B.2 and B.3

Proof of Lemma B.1. First of all, f is well-defined since both $x(t)$ and $y(t)$ are strictly increasing functions of t. Note that

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \phi(\alpha - t) - \phi(-\alpha - t), \quad \frac{\mathrm{d}y}{\mathrm{d}t} = 2t\left[\Phi(\alpha - t) - \Phi(-\alpha - t)\right].$$

Applying the chain rule gives

$$\begin{aligned} f'(t) = \frac{\mathrm{d}y}{\mathrm{d}t} \Big/ \frac{\mathrm{d}x}{\mathrm{d}t} &= \frac{2t\left[\Phi(\alpha - t) - \Phi(-\alpha - t)\right]}{\phi(\alpha - t) - \phi(-\alpha - t)} \\ &= \frac{2t\int_{-\alpha-t}^{\alpha-t} \mathrm{e}^{-\frac{u^2}{2}}\mathrm{d}u}{\mathrm{e}^{-\frac{(\alpha-t)^2}{2}} - \mathrm{e}^{-\frac{(-\alpha-t)^2}{2}}} = \frac{2t\int_{-\alpha}^{\alpha} \mathrm{e}^{-\frac{(u-t)^2}{2}}\mathrm{d}u}{\mathrm{e}^{-\frac{\alpha^2+t^2-2\alpha t}{2}} - \mathrm{e}^{-\frac{\alpha^2+t^2+2\alpha t}{2}}} \\ &= \frac{2t\mathrm{e}^{\frac{\alpha^2}{2}}\int_{-\alpha}^{\alpha} \mathrm{e}^{-\frac{u^2}{2}}\mathrm{e}^{tu}\mathrm{d}u}{\mathrm{e}^{\alpha t} - \mathrm{e}^{-\alpha t}} = \frac{2\mathrm{e}^{\frac{\alpha^2}{2}}\int_0^{\alpha} \mathrm{e}^{-\frac{u^2}{2}}(\mathrm{e}^{tu} + \mathrm{e}^{-tu})\mathrm{d}u}{\int_0^{\alpha} \mathrm{e}^{tu} + \mathrm{e}^{-tu}\mathrm{d}u}. \end{aligned}$$

Since $x(t)$ is strictly increasing in t, we see that $f''(t) \le 0$ is equivalent to saying that the function

$$g(t) := \frac{\int_0^{\alpha} \mathrm{e}^{-\frac{u^2}{2}}(\mathrm{e}^{tu} + \mathrm{e}^{-tu})\mathrm{d}u}{\int_0^{\alpha} \mathrm{e}^{tu} + \mathrm{e}^{-tu}\mathrm{d}u} \equiv \frac{\int_0^{\alpha} \mathrm{e}^{-\frac{u^2}{2}}\cosh(tu)\mathrm{d}u}{\int_0^{\alpha} \cosh(tu)\mathrm{d}u}$$

is decreasing in t. Hence, it suffices to show that

$$g'(t) = \frac{\int_0^{\alpha} \mathrm{e}^{-\frac{u^2}{2}} u\sinh(tu)\mathrm{d}u\int_0^{\alpha}\cosh(tv)\mathrm{d}v - \int_0^{\alpha} \mathrm{e}^{-\frac{u^2}{2}}\cosh(tu)\mathrm{d}u\int_0^{\alpha} v\sinh(tv)\mathrm{d}v}{\left(\int_0^{\alpha}\cosh(tv)\mathrm{d}v\right)^2} \le 0.$$

Observe that the numerator is equal to

$$\begin{aligned} &\int_0^{\alpha}\int_0^{\alpha} \mathrm{e}^{-\frac{u^2}{2}} u\sinh(tu)\cosh(tv)\mathrm{d}u\mathrm{d}v - \int_0^{\alpha}\int_0^{\alpha} \mathrm{e}^{-\frac{u^2}{2}} v\cosh(tu)\sinh(tv)\mathrm{d}u\mathrm{d}v \\ &\qquad = \int_0^{\alpha}\int_0^{\alpha} \mathrm{e}^{-\frac{u^2}{2}}\left(u\sinh(tu)\cosh(tv) - v\cosh(tu)\sinh(tv)\right)\mathrm{d}u\mathrm{d}v \\ &\qquad \overset{u\leftrightarrow v}{=} \int_0^{\alpha}\int_0^{\alpha} \mathrm{e}^{-\frac{v^2}{2}}\left(v\sinh(tv)\cosh(tu) - u\cosh(tv)\sinh(tu)\right)\mathrm{d}v\mathrm{d}u \\ &\qquad = \frac{1}{2}\int_0^{\alpha}\int_0^{\alpha} (\mathrm{e}^{-\frac{u^2}{2}} - \mathrm{e}^{-\frac{v^2}{2}})\left(u\sinh(tu)\cosh(tv) - v\cosh(tu)\sinh(tv)\right)\mathrm{d}u\mathrm{d}v. \end{aligned}$$

Then it is sufficient to show that

$$(\mathrm{e}^{-\frac{u^2}{2}} - \mathrm{e}^{-\frac{v^2}{2}})\left(u\sinh(tu)\cosh(tv) - v\cosh(tu)\sinh(tv)\right) \le 0$$

for all $u, v, t \ge 0$. To see this, suppose $u \ge v$ without loss of generality so that $\mathrm{e}^{-\frac{u^2}{2}} - \mathrm{e}^{-\frac{v^2}{2}} \le 0$ and

$$\begin{aligned} u\sinh(tu)\cosh(tv) - v\cosh(tu)\sinh(tv) &\ge v(\sinh(tu)\cosh(tv) - \cosh(tu)\sinh(tv)) \\ &= v\sinh(tu - tv) \ge 0. \end{aligned}$$

This analysis further reveals that $f''(t) < 0$ for $t > 0$. Hence, f is strictly concave. □

To prove the other two lemmas, we collect some useful facts about (B.5). This equation has (a) one positive root for $\delta \geq 1$ or $\delta < 1, \epsilon = \epsilon^\star$, (b) two positive roots for $\delta < 1$ and $\epsilon < \epsilon^\star$, and (c) no positive root if $\delta < 1$ and $\epsilon > \epsilon^\star$. In the case of (a) and (b), call $t(\epsilon, \delta)$ the positive root of (B.5) (choose the larger one if there are two). Then $t(\epsilon, \delta)$ is a decreasing function of ϵ. In particular, $t(\epsilon, \delta) \to \infty$ as $\epsilon \to 0$. In addition, $2(1-\epsilon)\left[(1+t^2)\Phi(-t) - t\phi(t)\right] + \epsilon(1+t^2) > \delta$ if $t > t(\epsilon, \delta)$.

Lemma B.4. *For any $0 < u < u^\star$, (B.6) has a unique root, denoted by $t^\star(u)$, in (α_0, ∞). In addition, $t^\star(u)$ strictly decreases as u increases, and it further obeys $0 < (1-u)/(1-2\Phi(-t^\star(u))) < 1$.*

Lemma B.5. *As a function of u,*

$$q^\star(u) = \frac{2(1-\epsilon)\Phi(-t^\star(u))}{2(1-\epsilon)\Phi(-t^\star(u)) + \epsilon u}$$

is strictly increasing on $(0, u^\star)$.

Proof of Lemma B.2. We first consider the regime: $\delta < 1, \epsilon > \epsilon^\star$. By (B.3), it is sufficient to show that $\mathrm{tpp}^\infty = 2(1-\epsilon')\Phi(-\alpha) + \epsilon' < u^\star$ under the constraints (B.4). From (B.4b) it follows that

$$\Phi(-\alpha) = \frac{1}{2}\,\mathbb{P}(|W| > \alpha) < \frac{\delta - \epsilon\epsilon'}{2(1-\epsilon\epsilon')},$$

which can be rearranged as

$$2(1-\epsilon')\Phi(-\alpha) + \epsilon' < \frac{(1-\epsilon')(\delta - \epsilon\epsilon')}{1-\epsilon\epsilon'} + \epsilon'.$$

The right-hand side is an increasing function of ϵ' because its derivative is equal to $(1-\epsilon)(1-\delta)/(1-\epsilon\epsilon')^2$ and is positive. Since the range of ϵ' is $(0, \epsilon^\star/\epsilon)$, we get

$$2(1-\epsilon')\Phi(-\alpha) + \epsilon' < \frac{(1-\epsilon^\star/\epsilon)(\delta - \epsilon\cdot\epsilon^\star/\epsilon)}{1-\epsilon\cdot\epsilon^\star/\epsilon} + \epsilon^\star/\epsilon = u^\star.$$

This bound $u^\star$ can be arbitrarily approached: let $\epsilon' = \epsilon^\star/\epsilon$ in the example given in Section 2.3; then set $\lambda = \sqrt{M}$ and take $M \to \infty$.

We turn our attention to the easier case where $\delta \geq 1$, or $\delta < 1$ and $\epsilon \leq \epsilon^\star$. By definition, the upper limit $u^\star = 1$ trivially holds. It remains to argue that tpp^∞ can be arbitrarily close to 1. To see this, set $\Pi^\star = M$ almost surely, and take the same limits as before: then $\mathrm{tpp}^\infty \to 1$.

□

Proof of Lemma B.3. We begin by first considering the boundary case:

$$\mathrm{tpp}^\infty = u. \tag{B.7}$$

In view of (B.3), we can write

$$\mathrm{fdp}^\infty = \frac{2(1-\epsilon)\Phi(-\alpha)}{2(1-\epsilon)\Phi(-\alpha) + \epsilon\mathrm{tpp}^\infty} = \frac{2(1-\epsilon)\Phi(-\alpha)}{2(1-\epsilon)\Phi(-\alpha) + \epsilon u}.$$

Therefore, a lower bound on fdp^∞ is equivalent to maximizing α under the constraints (B.4) and (B.7).

Recall $\mathbb{E}\,\eta_\alpha(W)^2 = 2(1+\alpha^2)\Phi(-\alpha) - 2\alpha\phi(\alpha)$. Then from (B.7) and (B.4a) we obtain a sandwiching expression for $1-\epsilon'$:

$$\frac{(1-\epsilon)\left[2(1+\alpha^2)\Phi(-\alpha) - 2\alpha\phi(\alpha)\right] + \epsilon(1+\alpha^2) - \delta}{\epsilon\left[(1+\alpha^2)(1-2\Phi(-\alpha)) + 2\alpha\phi(\alpha)\right]} < 1-\epsilon' = \frac{1-u}{1-2\Phi(-\alpha)},$$

which implies

$$\frac{(1-\epsilon)\left[2(1+\alpha^2)\Phi(-\alpha) - 2\alpha\phi(\alpha)\right] + \epsilon(1+\alpha^2) - \delta}{\epsilon\left[(1+\alpha^2)(1-2\Phi(-\alpha)) + 2\alpha\phi(\alpha)\right]} - \frac{1-u}{1-2\Phi(-\alpha)} < 0.$$

The left-hand side of this display tends to $1-(1-u) = u > 0$ as $\alpha \to \infty$, and takes on the value 0 if $\alpha = t^\star(u)$. Hence, by the uniqueness of $t^\star(u)$ provided by Lemma B.4, we get $\alpha < t^\star(u)$. In conclusion,

$$\mathrm{fdp}^\infty = \frac{2(1-\epsilon)\Phi(-\alpha)}{2(1-\epsilon)\Phi(-\alpha) + \epsilon u} > \frac{2(1-\epsilon)\Phi(-t^\star(u))}{2(1-\epsilon)\Phi(-t^\star(u)) + \epsilon u} = q^\star(u). \tag{B.8}$$

It is easy to see that fdp^∞ can be arbitrarily close to $q^\star(u)$.

To finish the proof, we proceed to consider the general case $\mathrm{tpp}^\infty = u' > u$. The previous discussion clearly remains valid, and hence (B.8) holds if u is replaced by u'; that is, we have

$$\mathrm{fdp}^\infty > q^\star(u').$$

By Lemma B.5, it follows from the monotonicity of $q^\star(\cdot)$ that $q^\star(u') > q^\star(u)$. Hence, $\mathrm{fdp}^\infty > q^\star(u)$, as desired.

□

B.2 Proofs of auxiliary lemmas

Proof of Lemma B.4. Set

$$\zeta := 1 - \frac{2(1-\epsilon)\left[(1+t^2)\Phi(-t) - t\phi(t)\right] + \epsilon(1+t^2) - \delta}{\epsilon\left[(1+t^2)(1-2\Phi(-t)) + 2t\phi(t)\right]}$$

or, equivalently,

$$2(1-\epsilon\zeta)\left[(1+t^2)\Phi(-t) - t\phi(t)\right] + \epsilon\zeta(1+t^2) = \delta. \tag{B.9}$$

As in Section B.1, we abuse notation a little and let $t(\zeta) = t(\epsilon\zeta, \delta)$ denote the (larger) positive root of (B.9). Then the discussion about (B.5) in Section B.1 shows that $t(\zeta)$ decreases as ζ increases. Note that in the case where $\delta < 1$ and $\epsilon > \epsilon^\star(\delta)$, the range of ζ in (B.9) is assumed to be $(0, \epsilon^\star/\epsilon)$, since otherwise (B.9) does not have a positive root (by convention, set $\epsilon^\star(\delta) = 1$ if $\delta > 1$).

Note that (B.6) is equivalent to

$$u = 1 - \frac{2(1-\epsilon)\left[(1+t^2)\Phi(-t) - t\phi(t)\right] + \epsilon(1+t^2) - \delta}{\epsilon\left[(1+t^2)(1-2\Phi(-t)) + 2t\phi(t)\right]/(1-2\Phi(-t))}. \tag{B.10}$$

Define

$$h(\zeta) = 1 - \frac{2(1-\epsilon)\left[(1+t(\zeta)^2)\Phi(-t(\zeta)) - t(\zeta)\phi(t(\zeta))\right] + \epsilon(1+t(\zeta)^2) - \delta}{\epsilon\left[(1+t(\zeta)^2)(1-2\Phi(-t(\zeta))) + 2t(\zeta)\phi(t(\zeta))\right]/(1-2\Phi(-t(\zeta)))}.$$

In view of (B.9) and (B.10), the proof of this lemma would be completed once we show the existence of ζ such that $h(\zeta) = u$. Now we prove this fact.

On the one hand, as $\zeta \searrow 0$, we see $t(\zeta) \nearrow \infty$. This leads to

$$h(\zeta) \to 0.$$

On the other hand, if $\zeta \nearrow \min\{1, \epsilon^\star/\epsilon\}$, then $t(\zeta)$ converges to $t^\star(u^\star) > \alpha_0$, which satisfies

$$2(1 - \min\{\epsilon, \epsilon^\star\}) \left[(1 + t^{\star 2})\Phi(-t^\star) - t^\star\phi(t^\star)\right] + \min\{\epsilon, \epsilon^\star\}(1 + t^{\star 2}) = \delta.$$

Consequently, we get

$$h(\zeta) \to u^\star.$$

Therefore, by the continuity of $h(\zeta)$, for any $u \in (0, u^\star)$ we can find $0 < \epsilon' < \min\{1, \epsilon^\star/\epsilon\}$ such that $h(\epsilon') = u$. Put $t^\star(u) = t(\epsilon')$. We have

$$\frac{1 - u}{1 - 2\Phi(-t^\star(u))} = 1 - \epsilon' < 1.$$

Last, to prove the uniqueness of $t^\star(u)$ and its monotonically decreasing dependence on u, it suffices to ensure that (a) $t(\zeta)$ is a decreasing function of ζ, and (b) $h(\zeta)$ is an increasing function of ζ. As seen above, (a) is true, and (b) is also true as can be seen from writing h as $h(\zeta) = 2(1 - \zeta)\Phi(-t(\zeta)) + \zeta$, which is an increasing function of ζ.

□

Proof of Lemma B.5. Write

$$q^\star(u) = \frac{2(1 - \epsilon)}{2(1 - \epsilon) + \epsilon u/\Phi(-t^\star(u))}.$$

This suggests that the lemma amounts to saying that $u/\Phi(-t^\star(u))$ is a decreasing function of u. From (B.10), we see that this function is equal to

$$\frac{1}{\Phi(-t^\star(u))} - \frac{(1 - 2\Phi(-t^\star(u))) \left\{2(1 - \epsilon) \left[(1 + (t^\star(u))^2)\Phi(-t^\star(u)) - t^\star(u)\phi(t^\star(u))\right] + \epsilon(1 + (t^\star(u))^2) - \delta\right\}}{\epsilon\Phi(-t^\star(u)) \left[(1 + (t^\star(u))^2)(1 - 2\Phi(-t^\star(u))) + 2t^\star(u)\phi(t^\star(u))\right]}.$$

With the proviso that $t^\star(u)$ is decreasing in u, it suffices to show that

$$\begin{aligned}
&\frac{1}{\Phi(-t)} - \frac{(1 - 2\Phi(-t)) \left\{2(1 - \epsilon) \left[(1 + t^2)\Phi(-t) - t\phi(t)\right] + \epsilon(1 + t^2) - \delta\right\}}{\epsilon\Phi(-t) \left[(1 + t^2)(1 - 2\Phi(-t)) + 2t\phi(t)\right]} \\
&\quad = \frac{\delta}{\epsilon} \cdot \underbrace{\frac{1 - 2\Phi(-t)}{\Phi(-t) \left[(1 + t^2)(1 - 2\Phi(-t)) + 2t\phi(t)\right]}}_{f_1(t)} - \frac{2}{\epsilon} \cdot \underbrace{\frac{(1 - 2\Phi(-t)) \left[(1 + t^2)\Phi(-t) - t\phi(t)\right]}{\Phi(-t) \left[(1 + t^2)(1 - 2\Phi(-t)) + 2t\phi(t)\right]}}_{f_2(t)} + 2
\end{aligned}$$

is an increasing function of $t > 0$. Simple calculations show that f_1 is increasing while f_2 is decreasing over $(0, \infty)$. This finishes the proof. □

C Proof of Theorem 1

With the results given in Appendices A and B in place, we are ready to characterize the optimal false/true positive rate trade-off. Up until now, the results hold for bounded λ, and we thus need to extend the results to arbitrarily large λ. It is intuitively easy to conceive that the support size of $\widehat{\boldsymbol{\beta}}$ will be small with a very large λ, resulting in low power. The following lemma, whose proof constitutes the subject of Section C.1, formalizes this point. In this section, $\sigma \geq 0$ may take on the value zero.

Lemma C.1. *For any $c > 0$, there exists λ' such that*

$$\sup_{\lambda > \lambda'} \frac{\#\{j : \widehat{\beta}_j(\lambda) \neq 0\}}{p} \leq c$$

holds with probability converging to one.

Assuming the conclusion of Lemma C.1, we prove claim (b) in Theorem 1 (noisy case), and then (a) (noiseless case). (c) is a simple consequence of (a) and (b), and (d) follows from Appendix B.

Case $\sigma > 0$. Let c be sufficiently small such that $q^\star(c/\epsilon) < 0.001$. Pick a large enough λ' such that Lemma C.1 holds. Then with probability tending to one, for all $\lambda > \lambda'$, we have

$$\mathrm{TPP}(\lambda) = \frac{T(\lambda)}{k \vee 1} \leq (1 + o_{\mathbb{P}}(1)) \frac{\#\{j : \widehat{\beta}_j(\lambda) \neq 0\}}{\epsilon p} \leq (1 + o_{\mathbb{P}}(1)) \frac{c}{\epsilon}.$$

On this event, we get

$$q^\star(\mathrm{TPP}(\lambda)) - 0.001 \leq q^\star(c/\epsilon + o_{\mathbb{P}}(1)) - 0.001 \leq 0,$$

which implies that

$$\bigcap_{\lambda > \lambda'} \Big\{ \mathrm{FDP}(\lambda) \geq q^\star\left(\mathrm{TPP}(\lambda)\right) - 0.001 \Big\} \tag{C.1}$$

holds with probability approaching one.

Now we turn to work on the range $[0.01, \lambda']$. By Lemma A.2, we get that $V(\lambda)/p$ (resp. $T(\lambda)/p$) converges in probability to $\mathrm{fd}^\infty(\lambda)$ (resp. $\mathrm{td}^\infty(\lambda)$) uniformly over $[0.01, \lambda']$. As a consequence,

$$\mathrm{FDP}(\lambda) = \frac{V(\lambda)}{\max\{V(\lambda) + T(\lambda), 1\}} \xrightarrow{\mathbb{P}} \frac{\mathrm{fd}^\infty(\lambda)}{\mathrm{fd}^\infty(\lambda) + \mathrm{td}^\infty(\lambda)} = \mathrm{fdp}^\infty(\lambda) \tag{C.2}$$

uniformly over $\lambda \in [0.01, \lambda']$. The same reasoning also justifies that

$$\mathrm{TPP}(\lambda) \xrightarrow{\mathbb{P}} \mathrm{tpp}^\infty(\lambda) \tag{C.3}$$

uniformly over $\lambda \in [0.01, \lambda']$. From Lemma B.3 it follows that

$$\mathrm{fdp}^\infty(\lambda) > q^\star(\mathrm{tpp}^\infty(\lambda)).$$

Hence, by the continuity of $q^\star(\cdot)$, combining (C.2) with (C.3) gives that

$$\mathrm{FDP}(\lambda) \geq q^\star(\mathrm{TPP}(\lambda)) - 0.001$$

holds simultaneously for all $\lambda \in [0.01, \lambda']$ with probability tending to one. This concludes the proof.

Case $\sigma = 0$. Fix λ and let $\sigma > 0$ be sufficiently small. We first prove that Lemma A.1 still holds for $\sigma = 0$ if α and τ are taken to be the limiting solution to (A.1) with $\sigma \to 0$, denoted by α' and τ'. Introduce $\widehat{\boldsymbol{\beta}}^\sigma$ to be the Lasso solution with data $\boldsymbol{y}^\sigma := \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{z} = \boldsymbol{y} + \boldsymbol{z}$, where $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$ is independent of $\boldsymbol{X}$ and $\boldsymbol{\beta}$. Our proof strategy is based on the approximate equivalence between $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}^\sigma$.

It is well known that the Lasso residuals $\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ are obtained by projecting the response $\boldsymbol{y}$ onto the polytope $\{\boldsymbol{r} : \|\boldsymbol{X}^\top \boldsymbol{r}\|_\infty \le \lambda\}$. The non-expansive property of projections onto convex sets gives

$$\left\|(\boldsymbol{y}^\sigma - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^\sigma) - (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})\right\| \le \|\boldsymbol{y}^\sigma - \boldsymbol{y}\| = \|\boldsymbol{z}\|.$$

If $\boldsymbol{P}(\cdot)$ is the projection onto the polytope, then $\boldsymbol{I} - \boldsymbol{P}$ is also non-expansive and, therefore,

$$\left\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}^\sigma - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right\| \le \|\boldsymbol{z}\|. \tag{C.4}$$

Hence, from Lemma A.3 and $\|\boldsymbol{z}\| = (1 + o_{\mathbb{P}}(1))\sigma\sqrt{n}$ it follows that, for any $c > 0$ and r_c depending on c,

$$\#\{1 \le j \le p : |\boldsymbol{X}_j^\top(\boldsymbol{y}^\sigma - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^\sigma - \boldsymbol{y} + \boldsymbol{X}\widehat{\boldsymbol{\beta}})| > 2r_c\sigma\} \le cp \tag{C.5}$$

holds with probability converging to one. Let $\boldsymbol{g}$ and $\boldsymbol{g}^\sigma$ be subgradients certifying the KKT conditions for $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}^\sigma$. From

$$\begin{aligned}\boldsymbol{X}_j^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) &= \lambda g_j,\\ \boldsymbol{X}_j^\top(\boldsymbol{y}^\sigma - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^\sigma) &= \lambda g_j^\sigma,\end{aligned}$$

we get a simple relationship:

$$\{j : |g_j| \ge 1 - a/2\} \setminus \{j : |g_j^\sigma| \ge 1 - a/2 - 2r_c\sigma/\lambda\} \subseteq \{j : |\boldsymbol{X}_j^\top(\boldsymbol{y}^\sigma - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^\sigma - \boldsymbol{y} + \boldsymbol{X}\widehat{\boldsymbol{\beta}})| > 2r_c\sigma\}.$$

Choose σ sufficiently small such that $2r_c\sigma/\lambda < a/2$, that is, $\sigma < a\lambda/(4r_c)$. Then

$$\{j : |g_j| \ge 1 - a/2\} \setminus \{j : |g_j^\sigma| \ge 1 - a\} \subseteq \{j : |\boldsymbol{X}_j^\top(\boldsymbol{y}^\sigma - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^\sigma - \boldsymbol{y} + \boldsymbol{X}\widehat{\boldsymbol{\beta}})| > 2r_c\sigma\}. \tag{C.6}$$

As earlier, denote by $\mathcal{S} = \mathrm{supp}(\widehat{\boldsymbol{\beta}})$ and $\mathcal{S}^\sigma = \mathrm{supp}(\widehat{\boldsymbol{\beta}}^\sigma)$. In addition, let $\mathcal{S}_v = \{j : |g_j| \ge 1 - v\}$ and similarly $\mathcal{S}_v^\sigma = \{j : |g_j^\sigma| \ge 1 - v\}$. Notice that we have dropped the dependence on λ since λ is fixed. Continuing, since $\mathcal{S} \subseteq \mathcal{S}_{\frac{a}{2}}$, from (C.6) we obtain

$$\mathcal{S} \setminus \mathcal{S}_a^\sigma \subseteq \{j : |\boldsymbol{X}_j^\top(\boldsymbol{y}^\sigma - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^\sigma - \boldsymbol{y} + \boldsymbol{X}\widehat{\boldsymbol{\beta}})| > 2r_c\sigma\}. \tag{C.7}$$

This suggests that we apply Proposition 3.6 of [3] that claims[6] the existence of positive constants $a_1 \in (0,1), a_2$, and a_3 such that with probability tending to one,

$$\sigma_{\min}(\boldsymbol{X}_{\mathcal{S}_{a_1}^\sigma \cup \mathcal{S}'}) \ge a_3 \tag{C.8}$$

for all $|\mathcal{S}'| \le a_2 p$. These constants also have positive limits a_1', a_2', a_3', respectively, as $\sigma \to 0$. We take $a < a_1', c < a_2'$ (we will specify a, c later) and sufficiently small σ in (C.7), and $\mathcal{S}' = \{j : |\boldsymbol{X}_j^\top(\boldsymbol{y}^\sigma - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^\sigma - \boldsymbol{y} + \boldsymbol{X}\widehat{\boldsymbol{\beta}})| > 2r_c\sigma\}$. Hence, on this event, (C.5), (C.7), and (C.8) together give

$$\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^\sigma\| = \left\|\boldsymbol{X}_{\mathcal{S}_a^\sigma \cup \mathcal{S}'}(\widehat{\boldsymbol{\beta}}_{\mathcal{S}_a^\sigma \cup \mathcal{S}'} - \widehat{\boldsymbol{\beta}}^\sigma_{\mathcal{S}_a^\sigma \cup \mathcal{S}'})\right\| \ge a_3\|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^\sigma\|$$

[6] Use a continuity argument to carry over the result of this proposition for finite t to ∞.

for sufficiently small σ, which together with (C.4) yields

$$\|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^{\sigma}\| \leq \frac{(1 + o_{\mathbb{P}}(1))\sigma\sqrt{n}}{a_3}. \tag{C.9}$$

Recall that the $\|\widehat{\boldsymbol{\beta}}\|_0$ is the number of nonzero entries in the vector $\widehat{\boldsymbol{\beta}}$. From (C.9), using the same argument outlined in (A.10), (A.11), and (A.12), we have

$$\|\widehat{\boldsymbol{\beta}}\|_0 \geq \|\widehat{\boldsymbol{\beta}}^{\sigma}\|_0 - c'p + o_{\mathbb{P}}(p) \tag{C.10}$$

for some constant $c' > 0$ that decreases to 0 as $\sigma/a_3 \to 0$.

We now develop a tight upper bound on $\|\widehat{\boldsymbol{\beta}}\|_0$. Making use of (C.7) gives

$$\|\widehat{\boldsymbol{\beta}}\|_0 \leq \|\widehat{\boldsymbol{\beta}}^{\sigma}\|_0 + \#\{j : 1 - a \leq |g_j^{\sigma}| < 1\} + cp.$$

As in (A.7), (3.21) of [3] implies that

$$\#\left\{j : (1-a) \leq |g_j^{\sigma}| < 1\right\}/p \xrightarrow{\mathbb{P}} \mathbb{P}\left((1-a)\alpha\tau \leq |\Pi + \tau W| < \alpha\tau\right).$$

Note that both α and τ depend on σ, and as $\sigma \to 0$, α and τ converge to, respectively, $\alpha' > 0$ and $\tau' > 0$. Hence, we get

$$\|\widehat{\boldsymbol{\beta}}\|_0 \leq \|\widehat{\boldsymbol{\beta}}^{\sigma}\|_0 + c''p + o_{\mathbb{P}}(p) \tag{C.11}$$

for some constant $c'' > 0$ that can be made arbitrarily small if $\sigma \to 0$ by first taking a and c sufficiently small.

With some obvious notation, a combination of (C.10) and (C.11) gives

$$|V - V^{\sigma}| \leq c'''p, \quad |T - T^{\sigma}| \leq c'''p, \tag{C.12}$$

for some constant $c''' = c'''_{\sigma} \to 0$ as $\sigma \to 0$. As $\sigma \to 0$, observe the convergence,

$$\mathrm{fd}^{\infty,\sigma} = 2(1-\epsilon)\Phi(-\alpha) \to 2(1-\epsilon)\Phi(-\alpha') = \mathrm{fd}^{\infty,0},$$

and

$$\mathrm{td}^{\infty,\sigma} = \epsilon\,\mathbb{P}(|\Pi^{\star} + \tau W| > \alpha\tau) \to \epsilon\,\mathbb{P}(|\Pi^{\star} + \tau' W| > \alpha'\tau') = \mathrm{td}^{\infty,0}.$$

By applying Lemma A.1 to V^{σ} and T^{σ} and making use of (C.12), the conclusions

$$\frac{V}{p} \xrightarrow{\mathbb{P}} \mathrm{fd}^{\infty,0} \quad \text{and} \quad \frac{T}{p} \xrightarrow{\mathbb{P}} \mathrm{td}^{\infty,0}$$

follow.

Finally, the results for some fixed λ can be carried over to a bounded interval $[0.01, \lambda']$ in exactly the same way as in the case where $\sigma > 0$. Indeed, the key ingredients, namely, Lemmas A.2 and A.4 still hold. To extend the results to $\lambda > \lambda'$, we resort to Lemma C.1.

For a fixed a prior Π, our arguments immediately give an instance-specific trade-off. Let $q^{\Pi}(\cdot; \delta, \sigma)$ be the function defined as

$$q^{\Pi}\left(\mathrm{tpp}^{\infty,\sigma}(\lambda); \delta, \sigma\right) = \mathrm{fdp}^{\infty,\sigma}(\lambda)$$

for all $\lambda > 0$. It is worth pointing out that the sparsity parameter ϵ is implied by Π and that q^{Π} depends on Π and σ only through Π/σ (if $\sigma = 0$, q^{Π} is invariant by rescaling). By definition, we always have

$$q^{\Pi}(u; \delta, \sigma) > q^{\star}(u)$$

for any u in the domain of q^{Π}. As is implied by the proof, it is impossible to have a series of instances Π such that $q^{\Pi}(u)$ converges to $q^{\star}(u)$ at *two* different points. Now, we state the instance-specific version of Theorem 1.

Theorem 3. *Fix $\delta \in (0, \infty)$ and assume the working hypothesis. In either the noiseless or noisy case, the event*

$$\bigcap_{\lambda \geq 0.01} \left\{ \mathrm{FDP}(\lambda) \geq q^{\Pi}\left(\mathrm{TPP}(\lambda)\right) - 0.001 \right\}$$

holds with probability tending to one.

C.1 Proof of Lemma C.1

Consider the KKT conditions restricted to $\mathcal{S}(\lambda)$:

$$\boldsymbol{X}_{\mathcal{S}(\lambda)}^{\top} \left(\boldsymbol{y} - \boldsymbol{X}_{\mathcal{S}(\lambda)} \widehat{\boldsymbol{\beta}}(\lambda) \right) = \lambda \boldsymbol{g}(\lambda).$$

Here we abuse the notation a bit by identifying both $\widehat{\boldsymbol{\beta}}(\lambda)$ and $\boldsymbol{g}(\lambda)$ as $|\mathcal{S}(\lambda)|$-dimensional vectors. As a consequence, we get

$$\widehat{\boldsymbol{\beta}}(\lambda) = (\boldsymbol{X}_{\mathcal{S}(\lambda)}^{\top} \boldsymbol{X}_{\mathcal{S}(\lambda)})^{-1} (\boldsymbol{X}_{\mathcal{S}(\lambda)}^{\top} \boldsymbol{y} - \lambda \boldsymbol{g}(\lambda)). \tag{C.13}$$

Notice that $\boldsymbol{X}_{\mathcal{S}(\lambda)}^{\top} \boldsymbol{X}_{\mathcal{S}(\lambda)}$ is invertible almost surely since the Lasso solution has at most n nonzero components for *generic* problems (see e.g. [28]). By definition, $\widehat{\boldsymbol{\beta}}(\lambda)$ obeys

$$\frac{1}{2} \|\boldsymbol{y} - \boldsymbol{X}_{\mathcal{S}(\lambda)} \widehat{\boldsymbol{\beta}}(\lambda)\|^2 + \lambda \|\widehat{\boldsymbol{\beta}}(\lambda)\|_1 \leq \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{X}_{\mathcal{S}(\lambda)} \cdot \boldsymbol{0}\|^2 + \lambda \|\boldsymbol{0}\|_1 = \frac{1}{2} \|\boldsymbol{y}\|^2. \tag{C.14}$$

Substituting (C.13) into (C.14) and applying the triangle inequality give

$$\frac{1}{2} \left(\|\lambda \boldsymbol{X}_{\mathcal{S}(\lambda)} (\boldsymbol{X}_{\mathcal{S}(\lambda)}^{\top} \boldsymbol{X}_{\mathcal{S}(\lambda)})^{-1} \boldsymbol{g}(\lambda)\| - \|\boldsymbol{y} - \boldsymbol{X}_{\mathcal{S}(\lambda)} (\boldsymbol{X}_{\mathcal{S}(\lambda)}^{\top} \boldsymbol{X}_{\mathcal{S}(\lambda)})^{-1} \boldsymbol{X}_{\mathcal{S}(\lambda)}^{\top} \boldsymbol{y}\| \right)^2 + \lambda \|\widehat{\boldsymbol{\beta}}(\lambda)\|_1 \leq \frac{1}{2} \|\boldsymbol{y}\|^2.$$

Since $\boldsymbol{I}_{|\mathcal{S}(\lambda)|} - \boldsymbol{X}_{\mathcal{S}(\lambda)} (\boldsymbol{X}_{\mathcal{S}(\lambda)}^{\top} \boldsymbol{X}_{\mathcal{S}(\lambda)})^{-1} \boldsymbol{X}_{\mathcal{S}(\lambda)}^{\top}$ is simply a projection, we get

$$\|\boldsymbol{y} - \boldsymbol{X}_{\mathcal{S}(\lambda)} (\boldsymbol{X}_{\mathcal{S}(\lambda)}^{\top} \boldsymbol{X}_{\mathcal{S}(\lambda)})^{-1} \boldsymbol{X}_{\mathcal{S}(\lambda)}^{\top} \boldsymbol{y}\| \leq \|\boldsymbol{y}\|.$$

Combining the last displays gives

$$\lambda \|\boldsymbol{X}_{\mathcal{S}(\lambda)} (\boldsymbol{X}_{\mathcal{S}(\lambda)}^{\top} \boldsymbol{X}_{\mathcal{S}(\lambda)})^{-1} \boldsymbol{g}(\lambda)\| \leq 2 \|\boldsymbol{y}\|, \tag{C.15}$$

which is our key estimate.

Since $\sigma_{\max}(\boldsymbol{X}) = (1 + o_{\mathbb{P}}(1))(1 + 1/\sqrt{\delta})$,

$$\lambda \|\boldsymbol{X}_{\mathcal{S}(\lambda)} (\boldsymbol{X}_{\mathcal{S}(\lambda)}^{\top} \boldsymbol{X}_{\mathcal{S}(\lambda)})^{-1} \boldsymbol{g}(\lambda)\| \geq (1 + o_{\mathbb{P}}(1)) \frac{\lambda \|\boldsymbol{g}(\lambda)\|}{1 + 1/\sqrt{\delta}} = (1 + o_{\mathbb{P}}(1)) \frac{\lambda \sqrt{|\mathcal{S}(\lambda)|}}{1 + 1/\sqrt{\delta}}. \tag{C.16}$$

As for the right-hand side of (C.15), the law of large numbers reveals that

$$2\|\boldsymbol{y}\| = 2\|\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{z}\| = (2 + o_{\mathbb{P}}(1))\sqrt{n(\|\boldsymbol{\beta}\|^2/n + \sigma^2)} = (2 + o_{\mathbb{P}}(1))\sqrt{p\,\mathbb{E}\,\Pi^2 + n\sigma^2}.$$

Combining the last two displays, it follows from (C.15) and (C.16) that

$$\|\widehat{\boldsymbol{\beta}}(\lambda)\|_0 \equiv |\mathcal{S}(\lambda)_\lambda| \le (4 + o_{\mathbb{P}}(1))\frac{(p\,\mathbb{E}\,\Pi^2 + n\sigma^2)(1 + 1/\sqrt{\delta})^2}{\lambda^2} = (1 + o_{\mathbb{P}}(1))\frac{Cp}{\lambda^2}$$

for some constant C. It is worth emphasizing that the term $o_{\mathbb{P}}(1)$ is independent of any $\lambda > 0$. Hence, we finish the proof by choosing any $\lambda' > \sqrt{C/c}$.

D Proof of Theorem 2

We propose two preparatory lemmas regarding the χ^2-distribution, which will be used in the proof of the theorem.

Lemma D.1. *For any positive integer d and $t \ge 0$, we have*

$$\mathbb{P}(\chi_d \ge \sqrt{d} + t) \le \mathrm{e}^{-t^2/2}.$$

Lemma D.2. *For any positive integer d and $t \ge 0$, we have*

$$\mathbb{P}(\chi_d^2 \le td) \le (\mathrm{e}t)^{\frac{d}{2}}.$$

The first lemma can be derived by the Gaussian concentration inequality, also known as the Borell's inequality. The second lemma has a simple proof:

$$\begin{aligned}\mathbb{P}(\chi_d^2 \le td) &= \int_0^{td} \frac{1}{2^{\frac{d}{2}}\Gamma(\frac{d}{2})} x^{\frac{d}{2}-1}\mathrm{e}^{-\frac{x}{2}}\mathrm{d}x \\ &\le \int_0^{td} \frac{1}{2^{\frac{d}{2}}\Gamma(\frac{d}{2})} x^{\frac{d}{2}-1}\mathrm{d}x = \frac{2(td)^{\frac{d}{2}}}{d2^{\frac{d}{2}}\Gamma(\frac{d}{2})}.\end{aligned}$$

Next, Stirling's formula gives

$$\mathbb{P}(\chi_d^2 \le td) \le \frac{2(td)^{\frac{d}{2}}}{d2^{\frac{d}{2}}\Gamma(\frac{d}{2})} \le \frac{2(td)^{\frac{d}{2}}}{d2^{\frac{d}{2}}\sqrt{\pi d}(\frac{d}{2})^{\frac{d}{2}}\mathrm{e}^{-\frac{d}{2}}} \le (\mathrm{e}t)^{\frac{d}{2}}.$$

Now, we turn to present the proof of Theorem 2. Denote by $\mathcal{S}$ a subset of $\{1, 2, \ldots, p\}$, and let $m_0 = |\mathcal{S} \cap \{j : \beta_j = 0\}|$ and $m_1 = |\mathcal{S} \cap \{j : \beta_j \ne 0\}|$. Certainly, both m_0 and m_1 depend on $\mathcal{S}$, but the dependency is often omitted for the sake of simplicity. As earlier, denote by $k = \#\{j : \beta_j \ne 0\}$, which obeys $k = (\epsilon + o_{\mathbb{P}}(1))p$. Write $\widehat{\boldsymbol{\beta}}_{\mathcal{S}}^{\mathrm{LS}}$ for the least-squares estimate obtained by regressing $\boldsymbol{y}$ onto $\boldsymbol{X}_{\mathcal{S}}$. Observe that (3.2) is equivalent to solving

$$\operatorname*{argmin}_{\mathcal{S}\subset\{1,\ldots,p\}} \|\boldsymbol{y} - \boldsymbol{X}_{\mathcal{S}}\widehat{\boldsymbol{\beta}}_{\mathcal{S}}^{\mathrm{L}S}\|^2 + \lambda|\mathcal{S}|. \tag{D.1}$$

As is clear from (3.2), we only need to focus on $\mathcal{S}$ with cardinality no more than $\min\{n, p\}$. Denote by $\widehat{\mathcal{S}}$ the solution to (D.1), and define $\widehat{m}_0$ and $\widehat{m}_1$ as before. To prove Theorem 2 it suffices to show the following: for arbitrary small $c > 0$, we can find λ and M sufficiently large such that (3.2) gives

$$\mathbb{P}(\widehat{m}_0 > 2ck \text{ or } \widehat{m}_1 \le (1-c)k) \to 0. \tag{D.2}$$

Assume this is true. Then from (D.2) we see that $\widehat{m}_0 \le 2ck$ and $\widehat{m}_1 > (1-c)k$ hold with probability tending to one. On this event, the TPP is

$$\frac{\widehat{m}_1}{k} > 1 - c,$$

and the FDP is

$$\frac{\widehat{m}_0}{\widehat{m}_0 + \widehat{m}_1} \le \frac{2ck}{2ck + (1-c)k} = \frac{2c}{1+c}.$$

Hence, we can have arbitrarily small FDP and almost full power by setting c arbitrarily small.

It remains to prove (D.2) with proper choices of λ and M. Since

$$\{\widehat{m}_0 > 2ck \text{ or } \widehat{m}_1 \le (1-c)k\} \subset \{\widehat{m}_0 + \widehat{m}_1 > (1+c)k\} \cup \{\widehat{m}_1 \le (1-c)k\},$$

we only need to prove

$$\mathbb{P}(\widehat{m}_1 \le (1-c)k) \to 0 \tag{D.3}$$

and

$$\mathbb{P}(\widehat{m}_0 + \widehat{m}_1 > (1+c)k) \to 0. \tag{D.4}$$

We first work on (D.3). Write

$$\boldsymbol{y} = \sum_{j \in \mathcal{S}, \beta_j = M} M\boldsymbol{X}_j + \sum_{j \in \overline{\mathcal{S}}, \beta_j = M} M\boldsymbol{X}_j + \boldsymbol{z}.$$

In this decomposition, the summand $\sum_{j \in \mathcal{S}, \beta_j = M} M\boldsymbol{X}_j$ is already in the span of $\boldsymbol{X}_{\mathcal{S}}$. This fact implies that the residual vector $\boldsymbol{y} - \boldsymbol{X}_{\mathcal{S}}\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}_{\mathcal{S}}$ is the same as the projection of $\sum_{j \in \overline{\mathcal{S}}, \beta_j = M} M\boldsymbol{X}_j + \boldsymbol{z}$ onto the orthogonal complement of $\boldsymbol{X}_{\mathcal{S}}$. Thanks to the independence among $\boldsymbol{\beta}, \boldsymbol{X}$ and $\boldsymbol{z}$, our discussion proceeds by conditioning on the random support set of $\boldsymbol{\beta}$. A crucial but simple observation is that the orthogonal complement of $\boldsymbol{X}_{\mathcal{S}}$ of dimension $n - m_0 - m_1$ has uniform orientation, independent of $\sum_{j \in \overline{\mathcal{S}}, \beta_j = M} M\boldsymbol{X}_j + \boldsymbol{z}$. From this fact it follows that

$$L(\mathcal{S}) := \|\boldsymbol{y} - \boldsymbol{X}_{\mathcal{S}}\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}_{\mathcal{S}}\|^2 + \lambda|\mathcal{S}| \overset{d}{=} \left(\sigma^2 + M^2(k - m_1)/n\right)\chi^2_{n-m_0-m_1} + \lambda(m_0 + m_1). \tag{D.5}$$

Call $E_{\mathcal{S},u}$ the event on which

$$L(\mathcal{S}) \le \sigma^2(n - k + 2u\sqrt{n-k} + u^2) + \lambda k$$

holds; here, $u > 0$ is a constant to be specified later. In the special case where $\mathcal{S} = \mathcal{T}$ and $\mathcal{T} \equiv \{j : \beta_j \neq 0\}$ is the true support, Lemma D.1 says that this event has probability bounded as

$$\begin{aligned}
\mathbb{P}(E_{\mathcal{T},u}) &= \mathbb{P}\left(\sigma^2\chi^2_{n-k} + \lambda k \le \sigma^2(n - k + 2u\sqrt{n-k} + u^2) + \lambda k\right) \\
&= \mathbb{P}\left(\chi^2_{n-k} \le n - k + 2u\sqrt{n-k} + u^2\right) \\
&\ge 1 - \mathrm{e}^{-\frac{u^2}{2}}.
\end{aligned} \tag{D.6}$$

By definition, $E_{\widehat{\mathcal{S}},u}$ is implied by $E_{\mathcal{T},u}$. Using this fact, we will show that $\widehat{m}_1$ is very close to k, thus validating (D.3). By making use of

$$\{\widehat{m}_1 \le (1-c)k\} \subset \{\widehat{m}_0 + \widehat{m}_1 \ge (k+n)/2\} \cup \{\widehat{m}_1 \le (1-c)k, \widehat{m}_0 + \widehat{m}_1 < (k+n)/2\},$$

we see that it suffices to establish that

$$\mathbb{P}(\widehat{m}_0 + \widehat{m}_1 \ge (k+n)/2) \to 0 \tag{D.7}$$

and

$$\mathbb{P}(\widehat{m}_1 \le (1-c)k, \widehat{m}_0 + \widehat{m}_1 < (k+n)/2) \to 0 \tag{D.8}$$

for some λ and sufficient large M. For (D.7), we have

$$\begin{aligned}
\mathbb{P}(\widehat{m}_0 + \widehat{m}_1 \ge (k+n)/2) &\le \mathbb{P}(\overline{E}_{\mathcal{T},u}) + \mathbb{P}(E_{\mathcal{T},u} \cap \{\widehat{m}_0 + \widehat{m}_1 \ge (k+n)/2\}) \\
&\le \mathbb{P}(\overline{E}_{\mathcal{T},u}) + \mathbb{P}(E_{\widehat{\mathcal{S}},u} \cap \{\widehat{m}_0 + \widehat{m}_1 \ge (k+n)/2\}) \\
&\le \mathbb{P}(\overline{E}_{\mathcal{T},u}) + \sum_{m_0+m_1 \ge (k+n)/2} \mathbb{P}(E_{\mathcal{S},u}) \\
&\le \mathrm{e}^{-\frac{u^2}{2}} + \sum_{m_0+m_1 \ge (k+n)/2} \mathbb{P}(E_{\mathcal{S},u}),
\end{aligned} \tag{D.9}$$

where the last step makes use of (D.6), and the summation is over all $\mathcal{S}$ such that $m_0(\mathcal{S}) + m_1(\mathcal{S}) \ge (k+n)/2$. Due to (D.5), the event $\mathbb{E}_{\mathcal{S},u}$ has the same probability as

$$\begin{aligned}
&\left(\sigma^2 + M^2(k-m_1)/n\right)\chi^2_{n-m_0-m_1} + \lambda(m_0+m_1) \le \sigma^2(n-k+2u\sqrt{n-k}+u^2) + \lambda k \\
&\iff \chi^2_{n-m_0-m_1} \le \frac{\sigma^2(n-k+2u\sqrt{n-k}+u^2) + \lambda k - \lambda(m_0+m_1)}{\sigma^2 + M^2(k-m_1)/n}.
\end{aligned} \tag{D.10}$$

Since $m_0 + m_1 \ge (k+n)/2$, we get

$$\sigma^2(n-k+2u\sqrt{n-k}+u^2) + \lambda k - \lambda(m_0+m_1) \le \sigma^2(n-k+2u\sqrt{n-k}+u^2) - \lambda(n-k)/2.$$

Requiring

$$\lambda > 2\sigma^2, \tag{D.11}$$

would yield $\sigma^2(n-k+2u\sqrt{n-k}+u^2) - \lambda(n-k)/2 < 0$ for sufficiently large n (depending on u) as $n-k \to \infty$. In this case, we have $\mathbb{P}(E_{\mathcal{S},u}) = 0$ whenever $m_0 + m_1 \ge (k+n)/2$. Thus, taking $u \to \infty$ in (D.9) establishes (D.7).

Now we turn to (D.8). Observe that

$$\begin{aligned}
&\mathbb{P}\left(\widehat{m}_1 \le (1-c)k \text{ and } \widehat{m}_0 + \widehat{m}_1 < (k+n)/2\right) \\
&\qquad \le \mathbb{P}(\overline{E}_{\mathcal{T},u}) + \mathbb{P}(E_{\mathcal{T},u} \cap \{\widehat{m}_1 \le (1-c)k \text{ and } \widehat{m}_0 + \widehat{m}_1 < (k+n)/2\}) \\
&\qquad \le \mathrm{e}^{-\frac{u^2}{2}} + \mathbb{P}(E_{\widehat{\mathcal{S}},u} \cap \{\widehat{m}_1 \le (1-c)k \text{ and } \widehat{m}_0 + \widehat{m}_1 < (k+n)/2\}) \\
&\qquad \le \mathrm{e}^{-\frac{u^2}{2}} + \sum_{m_0+m_1 < \frac{k+n}{2}, m_1 \le (1-c)k} \mathbb{P}(E_{\mathcal{S},u}).
\end{aligned} \tag{D.12}$$

For $m_0+m_1 < (k+n)/2$ and $m_1 \le (1-c)k$, notice that $n-m_0-m_1 > (n-k)/2 = (\delta-\epsilon+o_{\mathbb{P}}(1))p/2$, and $M^2(k-m_1)/n \ge cM^2k/n \sim (c\epsilon/\delta + o_{\mathbb{P}}(1))M^2$. Let $t_0 > 0$ be a constant obeying

$$\frac{\delta-\epsilon}{5}(1+\log t_0) + \log 2 < -1,$$

then choose M sufficiently large such that

$$\frac{2\sigma^2(\delta-\epsilon)+2\lambda\epsilon}{(\sigma^2+c\epsilon M^2/\delta)(\delta-\epsilon)} < t_0. \tag{D.13}$$

This gives

$$\frac{\sigma^2(n-k+2u\sqrt{n-k}+u^2)+\lambda k-\lambda(m_0+m_1)}{(\sigma^2+M^2(k-m_1)/n)(n-m_0-m_1)} < t_0$$

for sufficiently large n. Continuing (D.12) and applying Lemma D.2, we get

$$\begin{aligned}
&\mathbb{P}\left(\widehat{m}_1 \le (1-c)k \text{ and } \widehat{m}_0+\widehat{m}_1 < (k+n)/2\right) \\
&\qquad \le \mathrm{e}^{-\frac{u^2}{2}} + \sum_{m_0+m_1<\frac{k+n}{2}, m_1\le(1-c)k} \mathbb{P}(\chi^2_{n-m_0-m_1} \le t_0(n-m_0-m_1)) \\
&\qquad \le \mathrm{e}^{-\frac{u^2}{2}} + \sum_{m_0+m_1<\frac{k+n}{2}, m_1\le(1-c)k} (\mathrm{e}t_0)^{\frac{n-m_0-m_1}{2}} \\
&\qquad \le \sum_{m_0+m_1<(k+n)/2, m_1\le(1-c)k} (\mathrm{e}t_0)^{\frac{(\delta-\epsilon)p}{5}} \\
&\qquad \le \mathrm{e}^{-\frac{u^2}{2}} + 2^p(\mathrm{e}t_0)^{\frac{(\delta-\epsilon)p}{5}} \\
&\qquad \le \mathrm{e}^{-\frac{u^2}{2}} + \mathrm{e}^{-p}.
\end{aligned} \tag{D.14}$$

Taking $u \to \infty$ proves (D.8).

Having established (D.3), we proceed to prove (D.4). By definition,

$$\begin{aligned}
\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 + \lambda\|\widehat{\boldsymbol{\beta}}\|_0 &= \|\boldsymbol{y}-\boldsymbol{X}_{\widehat{\mathcal{S}}}\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}_{\widehat{\mathcal{S}}}\|^2 + \lambda\|\boldsymbol{\beta}^{\mathrm{LS}}_{\widehat{\mathcal{S}}}\|_0 \\
&\le \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_0 \\
&= \|\boldsymbol{z}\|^2 + \lambda k.
\end{aligned}$$

If

$$\lambda > \frac{\sigma^2\delta}{c\epsilon}, \tag{D.15}$$

then

$$\widehat{m}_0+\widehat{m}_1 \le \frac{\|\boldsymbol{z}\|^2}{\lambda} + k = (1+o_{\mathbb{P}}(1))\frac{\sigma^2 n}{\lambda} + k \le (1+c)k$$

holds with probability tending to one, whence (D.4).

To recapitulate, selecting λ obeying (D.11) and (D.15), and M sufficiently large such that (D.13) holds, imply that (D.2) holds. The proof of the theorem is complete.